International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Impact of EXplainable AI on Trust Evolution with AI Error Severity: Comparing Similar Instances and Saliency Map in a Baggage Screening Task

Yixuan Wang, Jieqiong Zhao, Yang Ba, Michelle V. Mancenido, Erin K. Chiou & Ross Maciejewski

To cite this article: Yixuan Wang, Jieqiong Zhao, Yang Ba, Michelle V. Mancenido, Erin K. Chiou & Ross Maciejewski (09 Oct 2025): Impact of EXplainable AI on Trust Evolution with AI Error Severity: Comparing Similar Instances and Saliency Map in a Baggage Screening Task, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2025.2564272

To link to this article: https://doi.org/10.1080/10447318.2025.2564272

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2025.2564272

Taylor & Francis
Taylor & Francis Group

Impact of EXplainable AI on Trust Evolution with AI Error Severity: Comparing Similar Instances and Saliency Map in a Baggage Screening Task

Yixuan Wang[a], Jieqiong Zhao[b], Yang Ba[a], Michelle V. Mancenido[c], Erin K. Chiou[d] and Ross Maciejewski[a]

[a]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA; [b]School of Computer and Cyber Sciences, Augusta University, Augusta, GA, USA; [c]School of Mathematical and Natural Sciences, Arizona State University, Glendale, AZ, USA; [d]The Polytechnic School, Arizona State University, Mesa, AZ, USA

ABSTRACT

Explainable Artificial Intelligence (XAI) can enhance trust in AI by offering cues that support human reasoning of AI behavior. Yet its effects on trust evolution remain unclear, especially when AI makes errors. This study examines how explanations of AI predictions influence human trust in AI-assisted decision-making under varying error severities. We tested two XAI visualizations, two AI error types, and three explanation strategies in simulated baggage screening tasks through an online study. Responses from 280 participants show that XAI representation significantly affects human compliance with AI during errors, while AI error type further shapes compliance after AI errors. AI Error type also impacts verification behaviors during AI errors, such as requesting explanations or ground truth. Moreover, strategies for conveying XAI influence perceived trust in AI, highlighting important implications for generalizing XAI effects beyond lab-based trust research.

KEYWORDS

Human-AI trust; trust evolution; trust calibration; explainable Artificial Intelligence (XAI)

1. Introduction

Artificial intelligence (AI) technologies have been increasingly adopted in various applications to enhance automation, especially in high-stakes and time-sensitive cases where the cost of system errors could result in catastrophic consequences, like security checks (Liang et al., 2019), autonomous driving (Rao & Frtunikj, 2018), and medical diagnosis (Park & Han, 2018). Sustained trust in AI is necessary to benefit from the AI's automation, but trust in AI shifts when AI inevitably commits errors or is asked to perform outside of its operating bounds (Hoffman et al., 2013; Yang et al., 2023). Therefore, establishing appropriate trust in such safety-critical scenarios, i.e., trusting AI when it is reliable, and detecting and intervening when AI is prone to errors, is crucial to balance AI's benefits with safety.

Because AI is imperfect and thus errors are inevitable, it is necessary to understand how human trust in imperfect AI evolves, how it is affected when AI make mistakes, and how it can be restored, particularly in high-stakes domains with minimal margin for error. Prior research has shown that when people observe AI errors, their trust in the AI and perceptions of its capabilities often decline, and trust in AI can be challenging to recover once eroded (Dietvorst et al., 2015; Yang et al., 2023) Unlike the evolution of trust in human-human relationships, in which trust builds gradually through intentionality, social exchanges, and mutual attributions (Lee & See, 2004), trust in AI tends to start high but rapidly declines following errors or perceived violations (Glikson & Woolley, 2020). AI systems that were not designed to show intentionality or other social cues further make trust recovery harder (de Visser et al., 2020), leading humans to be less forgiving of AI errors compared to those made by human counterparts.

CONTACT Ross Maciejewski rmacieje@asu.edu School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2025.2564272.

EXplainable AI (XAI) has been proposed as a social-like mechanism that bridges the gap between "black box" outputs of AI models and human comprehension of these outputs (Cramer et al., 2008; Islam et al., 2021). By providing transparency of AI models and reasoning of AI decisions, XAI aims to make AI processes more accessible and understandable to people, thereby enabling trust in its outputs and effective joint decision-making. Additionally, previous studies have found that providing AI transparency could facilitate trust recovery (Rebensky et al., 2021). Various XAI techniques have been developed to support human understanding of AI behavior (Doran et al., 2017), such as feature contribution (Ribeiro et al., 2016), similar instances (Cai et al., 2019), and counterfactuals (Wachter et al., 2018), and they each "explain" AI in fundamentally different ways. Empirical studies on the impact of XAI on human-AI collaboration have produced mixed conclusions: some research suggests that XAI can influence human behavior, calibrate trust in AI, and improve overall human-AI performance (Lai & Tan, 2019; Yang et al., 2020); while others have found no significant benefits (Bansal et al., 2021; Sivaraman et al., 2023). These discrepancies may stem from variations in explanation contexts and differences in target user groups: different tasks or user groups might better align with one XAI approach over another (Gerlings et al., 2022; Kim et al., 2023). Additionally, findings from social science research illustrate that the effectiveness of an explanation depends on the interplay between the *explainer* and *explainee* (Mayr et al., 2019; Tomsett et al., 2020), and an explanation is more effective when it is needed by the *explainee* (Miller, 2019).

These findings highlighted the need for a more comprehensive understanding of how XAI functions in diverse settings, such as *what type* of XAI works best in which contexts and *how* to best deliver the XAI information, rather than simply focusing on XAI demonstration in well-constrained task environments. People can continuously adjust their level of trust in the system as they interact with AI over time, and this process can be further improved when XAI is involved. XAI serves as a mechanism for ongoing verification, allowing people to refine their mental models of AI's performance and reliability over time (Kulesza et al., 2013). Thus, it is important to investigate how human trust in AI evolves over time with the presence of XAI, and how diverse XAI approaches and their delivery strategies affect trust before, during, and after AI errors are observed in a high-stakes, time-sensitive task. Our study investigated the role of XAI presentation in trust development over time, with the occurrences of trust violations triggered by people observing AI errors during baggage screening. Specifically, our study aims to explore three research questions:

RQ1. How do different XAI representations influence trust formation, violation, and repair in AI-assisted baggage screening?

RQ2. How do different types of AI errors influence people's behavior and trust adaptation in the presence of XAI?

RQ3. How does the timing and availability of XAI explanations affect trust recalibration and reliance on AI after errors?

To address these questions, two types of XAI representations were tested as interventions for trust evolution, namely, the Saliency Map and Similar Instances, representing feature-based and example-based explanations for images respectively. Saliency Maps highlight features critical to a prediction, aiding causal attribution (Chou et al., 2022), while Similar Instances offer analogical reasoning through context-rich comparisons (Cai et al., 2019). These two techniques were selected because although they differ in how they align with human cognitive preferences in context-dependent scenarios, the cognitive load they impose, and the distinct mechanisms they use to achieve interoperability, both have been shown to effectively interpret the rationale behind specific outputs and contribute to the formation of human mental models (Cai et al., 2019; Guidotti et al., 2020; Petsiuk et al., 2021). Furthermore, both techniques have demonstrated potential in calibrating human trust in AI (Wang & Yin, 2022; Yang et al., 2020). Yet, the effects of both XAI techniques on how trust evolves over time in the context of AI errors remains understudied. Beyond explanation techniques, we also draw insights from human factors research by investigating how the availability of XAI presentation – whether continuously provided for all tasks, selectively offered for difficult tasks, or made available only upon request – impacts trust. These strategies reflect the notion that explanations are social (Chiou & Lee, 2023; de Visser et al., 2020; Miller, 2019) and therefore their efficacy depends on their responsivity to the human receiver's cognitive state and task context.

In addition, AI error types have been shown to impact people's perception of AI performance (Tolmeijer et al., 2021), their trust in AI, and their willingness to forgive its failures (Baughan et al., 2023). For example, in automated signal detection systems, false alarms and missed signals by the AI have been shown to diversely impact performance, trust, and monitoring behavior (Ferraro & Mouloua, 2021; Moray, 2003). Further, there is evidence that competency-based errors, e.g., malfunctions; and integrity-based errors, e.g., misaligned goals; require distinct strategies to repair trust (Marinaccio et al., 2015; Quinn et al., 2017). Therefore, in this work, we also explored the effect of different types of AI errors on the evolution of human trust in AI. We focused on two common AI detection errors: False Alarms and Missed Signals. These errors can present varying levels of perceived risk or severity to human collaborators, with Missed Signals potentially causing more severe consequences than False Alarms in safety-critical tasks.

To simulate a high-stakes and time-sensitive environment, we developed an X-ray baggage screening test bed based on airport security checkpoints workflow (Hartnett et al., 2022). The test bed follows current officer screening workflows, wherein an AI system screens baggage for threats, while the human supervisor make the final decision to either flag baggage for secondary inspection or clear it. After some exposure to a series of correct AI decisions designed to initiate the trust formation phase (i.e., initial high level of trust), we then deliberately presented consecutive incorrect AI decisions to induce a trust violation. Because participants were not informed of the ground truth, we selected tasks that were relatively easy to verify upon closer inspection, such that a non-expert (someone who is not deliberately trained in X-ray threat detection) could be expected to reasonably recognize, with some effort, if the AI produced errors.

Our study employed a factorial design with variables *XAI representation*, *strategy of conveying XAI*, and *type of AI error* to investigate how these factors together affect emergent human behavior, joint human-AI decision accuracy, and perceived trust in AI. Our contributions include: (1) using carefully designed tasks and test beds that emulate a high-stakes and time-sensitive scenario (i.e., baggage screening), we demonstrate how different XAI approaches targeting distinct human cognitive tendencies influence human trust in AI over time; (2) we examine the impact of error typology in object detection-based AI on human-AI trust; and (3) we offer design recommendations for implementing XAI in high-stakes and time-sensitive scenario settings with a low tolerance for errors.

2. Related work

Human-AI collaboration spans a broad field of study, but two key dimensions shape our work: (1) the impact of XAI on human-AI collaboration dynamics and (2) the evolution of human trust in AI over time. These areas frame the research gap our work addresses.

2.1. XAI in human-AI collaboration

Many XAI methods have been proposed to improve AI transparency by explaining the reasoning processes of AI decisions to human operators (Cramer et al., 2008; Islam et al., 2021). The two most commonly studied XAI categories in HCI area are feature-based explanations, such as feature importance (Ribeiro et al., 2016) and saliency map (Mundhenk et al., 2020), and example-based explanations, including nearest neighbors (Cai et al., 2019) and counterfactuals (Zytek et al., 2022; Wachter et al., 2018). Feature-based explanations, like Saliency Map, allow people to understand what the model sees as important for a prediction. This could appeal to an operator's need for causal attribution by providing a mechanism to identify the root cause of a mistake or to resolve uncertainties about the AI's decision process. In contrast, example-based techniques can provide tangible, context-rich comparisons to help people relate current task predictions to past outcomes. For instance, Similar Instances (i.e., nearest neighbors) explain an AI's decision by presenting people with past examples that are most similar to the current case. Instead of providing abstract rules or technical details, this method shows how the AI has handled comparable situations, making the reasoning process more concrete and relatable. This approach would appeal to an operator's analogical reasoning, which is often used in complex reasoning tasks where contextual understanding is required.

Many researchers have investigated the effects of these XAI technologies on human-AI joint decision-making (Chromik et al., 2021; Ooge et al., 2022; Tomsett et al., 2020; Zhang et al., 2020), both in terms of collaborative performance and human perception of AI, yet the findings are mixed. On the one hand, several studies reported the benefits of XAI on human-AI collaborative performance. For instance, Yang et al. (2020) found that adding visual explanations in human-AI collaboration environments yielded better decisions compared to people and AI working separately; Leichtmann et al. (2024, 2023) investigated the combined use of feature-based and example-based XAI, observing that this pairing enhanced decision accuracy, promoted appropriate human trust in AI, and highlighted the potential of explanations to signal AI errors effectively. Other work showed explanations can increase human confidence in AI (Sivaraman et al., 2023), improve human performance (Lai & Tan, 2019), raise human initial trust in AI (Ooge et al., 2022), and support better trust calibration in high-stakes settings (Tomsett et al., 2020). On the other hand, several studies found that providing XAI information did not guarantee improvements in human-AI collaboration performance (Alufaisan et al., 2021; Bansal et al., 2021; Buçinca et al., 2020; Liu et al., 2021; Wang & Yin, 2021) nor did XAI information affect human behaviors (Sivaraman et al., 2023). Explanations may not always contribute to better trust alignment or performance (Nourani et al., 2019; Zhang et al., 2020). These discrepancies in effects of XAI on human-AI collaboration can be attributed to myriad factors and their interaction effects, including explanation quality (Kunkel et al., 2019; Morrison et al., 2024; Poursabzi-Sangdeh et al., 2021), fidelity (Kulesza et al., 2013), representation (Cai et al., 2019; Yang et al., 2020), interactivity (Cheng et al., 2019; Liu et al., 2021), explanation meaningfulness (Nourani et al., 2019), AI accuracy (Papenmeier et al., 2022), and domain knowledge of workers (Dikmen & Burns, 2022; Nourani et al., 2020b; Szymanski et al., 2022; Wang & Yin, 2021; 2022).

Recent work has also focused on comparing the effectiveness of different XAI approaches in AI-assisted decision-making. For example, Chen et al. (2023) found that presenting example-based explanations can significantly improve decision performance compared to conditions without XAI, while feature-based explanations cannot; Humer et al. (2022) showed that example-based explanations are better for improving complementary human-AI decision performance than feature-based explanations. In contrast, Wang and Yin (2022) found that feature contribution explanation appears to enhance people's subjective understanding of AI, compared to example-based explanation. These mixed outcomes may stem from mismatches between explanation types and task contexts (Mayr et al., 2019; Prinster et al., 2024; Tomsett et al., 2020), the timing of when an explanation is shown to close knowledge gaps as needed (Miller, 2019), and people's varying responses to error typology (Ferraro & Mouloua, 2021). These considerations underscore the importance of carefully aligning explanation strategies with user needs and task demands when deploying XAI in human-AI collaborative systems.

2.2. Human trust in AI over time

As people interact with AI, their trust in the AI directly affects their willingness to adopt, rely on, and collaborate effectively with AI systems, especially in high-stakes or decision-critical contexts. A few key concepts of human trust in AI include their propensity to trust, their perceived trustworthiness of the AI, and their trusting behavior (Mayer et al., 1995; Schlicker et al., 2025). While propensity to trust reflects a relatively stable individual tendency, situational trust can vary substantially depending on contextual factors. Therefore, studying human–AI trust in specific situations and across time provides important insights into the dynamic nature of trust formation and calibration in human–AI interaction. Many studies on human-AI trust focus instead on measuring trusting behavior in specific tasks, which serve as observable indicators of trust, e.g., reliance, compliance (de Visser et al., 2020; Eckhardt et al., 2024; Vereschak et al., 2021). In addition, perceived trustworthiness is another commonly used metric in human-AI research. It is inherently situational and subjective, reflecting people's evaluations of the AI's competence, integrity, and intentions, as well as their belief in the AI's ability to support their task goals (Baer & Colquitt, 2018; Chiou & Lee, 2023). Moreover, various factors have been shown to potentially influence human-AI trust (Hoff & Bashir, 2015), including individual differences such as cultural background (Chien et al., 2020), familiarity with automated systems (Dasgupta et al., 2017), as well as contextual elements like task difficulty and workload (Hoff & Bashir, 2015), and AI system attributes such as AI performance (Kay et al., 2015) and explainability (Yang et al., 2020).

Beyond investigating the various factors and antecedents of human trust in AI systems, a growing array of studies have been conducted to explore human initial trust formation and trust evolution over time. For example, Cabiddu et al. (2022) proposed the factors that affect human initial trust in AI and trust over time separately. Holliday et al. (2016) performed a case study to compare how trust in AI changes with and without explanations. Several studies (Kahr et al., 2023; Yu et al., 2016; 2017) observed that trust levels are related to AI performance, and that trust tends to stabilize over time. Previous work also revealed that first impressions of AI, as well as the order of presenting AI's correct decisions, are pivotal for the development of trust in AI. Nourani et al. (2020b) found that people with domain knowledge are more sensitive to the first impression of AI compared to laypeople, where domain experts exhibited higher variance in trust over time if they had a positive first impression of AI. Conversely, laypeople usually overestimated AI performance and overtrusted AI (Nourani et al., 2020b). Later, Nourani et al. (2022) explored how initial impressions shape people's mental models of AI capabilities. They found that individuals initially exposed to correct AI predictions were more likely to over-rely on AI, whereas those who were initially exposed to incorrect AI predictions were more likely to disregard AI suggestions and rely more on their own judgment. Although first impression matters, we know from previous research that observing AI mistakes negatively impacts human trust, and that trust in AI can be difficult to restore once dropped (Dietvorst et al., 2015; Hald et al., 2021; Hoffman et al., 2013; Yang et al., 2023). Other research shows that various types of AI errors not only lead to differing perceptions of AI accuracy (Tolmeijer et al., 2021) but also impact the degree of trust placed in the AI and people's willingness to "forgive" the AI (Baughan et al., 2023).

Algorithm aversion is defined as people's reluctance to rely on algorithmic recommendations after observing errors, despite the algorithm typically outperforming human judgment (Dietvorst et al., 2015). Various methods have been proposed for overcoming algorithm aversion (Burton et al., 2020; Mahmud et al., 2022) and repairing human trust in AI after trust drops (De Visser et al., 2018). Tolmeijer et al. (2021) found that initial experience significantly influences trust levels, with early system accuracy leading to higher trust. Conversely, Kahr et al. (2024) suggest that while early errors do decrease trust and reliance (i.e., adoption of AI advice), trust can be rapidly restored, and reliance may not significantly drop after late errors. These results indicate that people may develop tolerance to case-by-case AI errors over time if the system consistently performs well (Kahr et al., 2024). Techniques such as enabling people to modify AI outputs (Dietvorst et al., 2018), informing them that the model has been updated to address errors Pareek et al. (2024), or demonstrating AI's capacity to learn from its mistakes (Reich et al., 2023), have shown potential for mitigating algorithm aversion and encouraging AI adoption in subsequent tasks. Moreover, efforts such as offering anthropomorphic apologies for AI failures (Kim & Song, 2021) and providing explanations for mistakes (Esterwood & Robert, 2021; Hald et al., 2021) attempt to improve the efficiency of trust repair. Although the effectiveness of trust repair approaches matters, Robinette et al. (2015) emphasized that the timing of implementing those methods is also crucial for the success of trust repair. The complex dynamics of human trust in AI pose significant challenges in understanding how trust in AI evolves over time, particularly in complex work environments where AI is susceptible to producing and contributing to various task errors (Hoffman et al., 2013; Hu et al., 2019; Yu et al., 2017). This underscores the importance of investigating how trust changes under varying types of AI errors, and whether specific interventions (e.g., XAI) can promote more calibrated and resilient human trust in AI systems.

2.3. Research gap

While studies have examined the impact of XAI on human trust in AI, very few have focused on evaluating the usefulness of XAI in high-stakes and time-sensitive scenarios. Though some previous work discovered that explanations of AI can effectively calibrate human-AI trust, especially for the feature-based and example-based XAI, there is limited understanding of how XAI shapes the trajectory of trust over time, particularly in scenarios where trust could decline when people notice AI errors. Since such errors are inevitable in real-world systems, prior studies have focused on trust repair interventions deployed immediately after errors occur, such as explanations, apologies, or denials (Esterwood & Robert, 2021; Hald et al., 2021; Kim & Song, 2021). However, detecting trust violations in real time can be difficult, making it even more challenging to effectively intervene in the process and apply appropriate methods to repair trust. Thus, our work takes a broader perspective. Rather than centering solely on trust repair, our goal was to understand how XAI influences the evolution of human trust in AI across the entire interaction, including both pre- and post-error phases, by continuously integrating XAI into the decision-making process. Prior research has shown the role of AI error types in shaping human trust, yet their effect on trust evolution in the presence of XAI remains overlooked. Because the effectiveness of explanations may depend on the interaction between the *explainer* (AI) and the *explainee* (human) (Mayr et al., 2019; Miller, 2019; Tomsett et al., 2020), it is important to investigate how different XAI delivery strategies, tailored to this interaction, affect how people's trust in AI evolves.

Besides, baggage screening task is distinct from other high-stakes tasks (e.g., medical diagnosis) because it involves rapid, repetitive visual search under time pressure, often with limited contextual information or feedback, low target prevalence, and high variability in visual appearances of threats (Schwaninger et al., 2005; Wolfe & Van Wert, 2010). While prior research in ergonomics has explored factors that influence automation-assisted baggage screening performance, such as the plausibility of visual cues for detected items (Chavaillaz et al., 2020), working environment for screeners (Latscha et al., 2024), and the type of detection errors (Huegli et al., 2025), human trust in automated baggage screening system and its role in joint human-automation performance remain underexplored, especially when explanations are provided for automated detection results. To address these gaps, our study explores how people's trust in AI evolves in the baggage screening task, examining how it is affected by different XAI approaches, strategies of conveying XAI, and the types of AI errors.

3. Methods

We conducted an online human-subject study on Prolific, where participants were asked to cooperate with an AI to perform baggage screening tasks. The study has been approved by Arizona State University's Institutional Review Board (IRB) and preregistered on OSF.[1]

3.1. Experimental design

3.1.1. Task validation

This work aims to explore the impact of XAI on trust evolution, particularly under scenarios where the consequences of errors are high. For this reason, we deployed an object detection task that has implications in security screening in public spaces – an X-ray baggage screening task with an automated threat detection system commonly deployed at airports (Hartnett et al., 2022; Hättenschwiler et al., 2018). In the U.S., transportation security officers (TSOs) are faced with the challenge of examining high volumes of luggage with varying levels of visual complexity (e.g., cluttered, objects rotated, and many possible threats, among others) within strict time limits, increasing the risk of missed detections. Despite technological advancements in imaging technology and embedded AI, TSOs are still required "trust but verify" AI outputs, given the low-risk tolerance in security screening work.

We consider the baggage screening task a plausible choice for engaging a general population of adults with at least a high school degree (which is the minimum education level required to be hired as a TSO), because it represents a common object detection scenario that is more intuitive than specialized tasks like monitoring autonomous driving systems or face matching. Participants were provided with tutorial materials detailing six target objects familiar from daily life, reducing the difficulty of the general baggage screening tasks. Further, participants performed baggage screening tasks without AI assistance in the pilot study and achieved a high detection accuracy of 91.3%. This result further confirms that a general population, after receiving basic tutorials, possesses the necessary skills to perform baggage screening effectively in our selected tasks. Please note that TSOs typically have a high school education level and historically experienced high turnover (Congress, 2020; Keller, 2019), meaning that many TSOs will have limited on-job experience. The foundations for this study and its applicability to real-world scenarios are aligned with checkpoint requirements (TSA, 2024) and have been developed with domain experts.

3.1.2. Task environment

Participants engaged in an AI-assisted decision-making process to detect illicit objects in baggage. The baggage screening image data used in the study was from SIXray (Miao et al., 2019), an object detection benchmark dataset that contains 1,059,231 images in total, consisting of both positive (with illicit object) and negative (without illicit object) X-ray baggage screening images. Among these, only 8,929 positive images are annotated with labels, covering five classes of illicit objects—Gun, Knife, Wrench, Pliers, and Scissors—excluding Hammer, despite it being part of the original six classes.[2] After cleaning the annotated data and removing images with annotation errors, we retained 8,908 valid positive images. From this cleaned set, we randomly sampled 9,256 images (7,756 positive, 1,500 negative) to train a class-balanced hierarchical refinement (CHR) (Miao et al., 2019), and evaluated its performance on another 1,252 images (1,152 positive, 100 negative) sampled from the remaining annotated images.

To select baggage screening images used in the study, we first manually categorized all AI-correct test images into two levels of task difficulty: easy and difficult. Tasks were labeled difficult if items in the baggage overlapped significantly or had substantial rotation; otherwise, they were considered easy. In addition, the difficulty level of selected images was further validated in the pilot study (Section 3.4). For each difficulty level, we randomly sampled 18 AI-correct baggage screening tasks (12 positive, 6 negative), resulting in 36 AI-correct baggage screening tasks in total for the study. To induce human trust violation by highlighting obvious AI failures, we selected three additional easy tasks for both the positive and negative cases. We deliberately manipulated their AI results to be incorrect, and positioned these tasks consecutively within the study (Section 3.1.3). To eliminate order bias, we randomized the order of 36 AI-correct tasks and randomized the order of three AI-error tasks separately.

For real-world baggage screening tasks, both screening accuracy and operational efficiency are crucial. However, there is often a trade-off between accuracy and efficiency. For example, if screeners are not confident in identifying illicit objects directly by AI or their own, they may need additional labor or time to manually examine the baggage to ensure accuracy. Thus, participants were first asked their agreement with AI detection results, then we gave them the option to open baggage for further inspection. Participants were informed that opening baggage was to confirm the existence of illicit objects in the baggage. It would return a 100% accurate result of all illicit objects in the baggage but would impose a 10-second delay before the next baggage. Unlike real-world screening environments, where separate staff typically perform the manual inspection by opening baggage flagged for containing or potentially containing illicit objects without interrupting the screener's flow, our protocol made the time cost of opening baggage fall directly on the participants. This design encouraged participants to weigh the trade-off between accuracy and efficiency, echoing high-stakes and time-sensitive settings. To make a final decision, participants should decide to either adopt the AI detection result directly, make the detection on their own, or consume extra time confirming the existence of actual illicit objects by opening the baggage. To emulate real-world baggage screening scenarios, participants were not informed of the ground truth or the AI correctness for each task during the study unless they chose to open the baggage for confirmation. In addition, we imposed a time limit for task completion to mimic the time pressure encountered in real baggage screening scenarios, without informing participants of the total number of tasks. As such, we expect that the goal for participants in our study is to achieve both high accuracy and efficiency, i.e., accurately inspecting as much baggage as possible within a time limit.

3.1.3. Trust manipulation

People's trust in AI can be determined by assessing their perception of AI decision performance (Cabiddu et al., 2022; Jacovi et al., 2021; Lee & See, 2004). For example, observing an AI's accurate

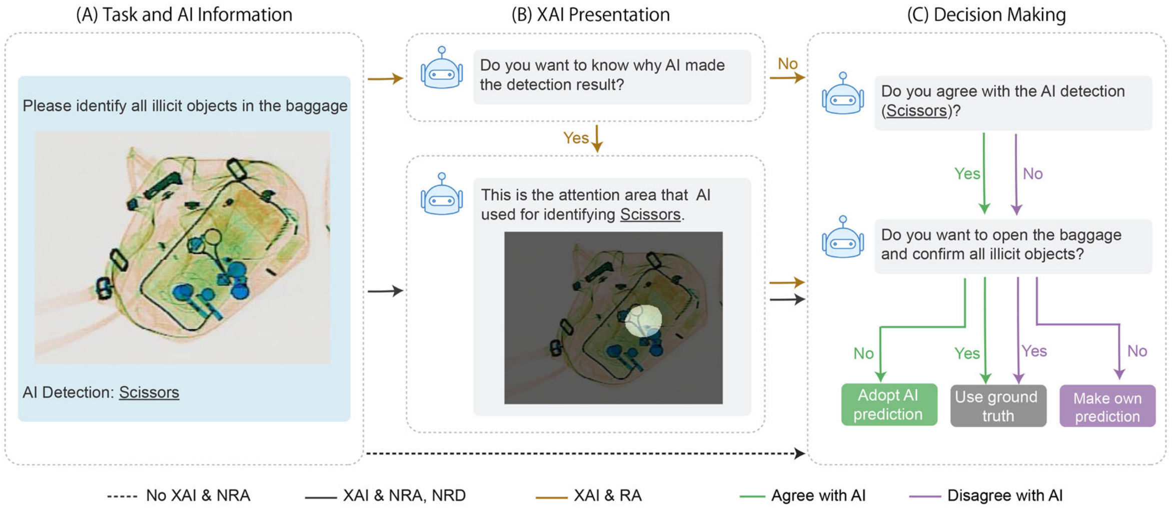

(A) Task and AI Information
Please identify all illicit objects in the baggage
AI Detection: Scissors
(B) XAI Presentation
Do you want to know why AI made the detection result?
Yes
This is the attention area that AI used for identifying Scissors.
No
(C) Decision Making
Do you agree with the AI detection (Scissors)?
Yes
No
Do you want to open the baggage and confirm all illicit objects?
No
Yes
Yes
No
Adopt AI prediction
Use ground truth
Make own prediction
No XAI & NRA
XAI & NRA, NRD
XAI & RA
Agree with AI
Disagree with AI

Figure 1. The interface and procedure of performing a baggage screening task with AI assistance in the study, all images on the interface can be zoomed in and out. This is an example of the Saliency Map condition.

detection could lead to high trust in the AI, while observing AI errors could lead to a trust drop. To explore these dynamics, our study manipulated human trust levels in an AI by controlling for the correctness of AI decisions. In particular, we structured the study into three trust phases: trust formation (McKnight et al., 1998; Pareek et al., 2024), trust violation (Lewicki & Bunker, 1996; Pareek et al., 2024), and trust repair (Lewicki & Bunker, 1996; Pareek et al., 2024), which were implemented as a within-subject factor (Figure 2).

3.1.3.1. Trust violation. In real-world human-AI collaboration scenarios, people inevitably encounter AI errors and become conscious of them, even without knowing the ground truth, which can quickly reduce their trust in the system (Dietvorst et al., 2015; Parasuraman & Riley, 1997; Yang et al., 2023). Therefore, we randomly selected easy baggage screening tasks for each type of AI error, manipulated the AI decisions to be incorrect, and presented them consecutively to participants to erode their trust in the AI. These tasks were deliberately chosen to be easy for the participants to do on their own (Supplementary Appendix A), allowing them to easily identify illicit objects themselves and recognize AI errors without receiving ground truth. Moreover, trust can decline after just a single perceived system error (Hald et al., 2021; Kahr et al., 2024), and continuous errors can further exacerbate this decline (Wang & Yin, 2023; Yu et al., 2017). Since this work focused on trust evolution and expected trust recovery, it was important to avoid solidifying participants' negative impressions with excessive AI errors. Therefore, we first examined the effect of a single error versus three consecutive errors on the deterioration of trust in AI and the subsequent process of trust recovery in a pilot study (Section 3.4). Based on the pilot study findings, we decided to expose participants to three consecutive AI errors to guarantee the "human-AI trust violation" condition in the formal study.

3.1.3.2. Trust formation and trust repair. Human trust may start low due to the initial unfamiliarity about the AI system but often grows as people gain more experience and confidence in the system (Parasuraman & Riley, 1997). In addition, a positive first impression (that is, satisfying initial expectations of system accuracy) was found to effectively promote human trust in AI systems (Tolmeijer et al., 2021; Yu et al., 2017). To establish high human trust in AI during the early stages of human-AI interaction and thus subsequently ensure a significant trust violation following observed AI errors, we presented 12 consecutive baggage screening tasks with correct AI detection results before Trust Violation, forming the Trust Formation phase. Furthermore, our work attempted to recover human trust by showing another 24 AI-correct tasks after three consecutive AI errors in the Trust Violation phase, forming the Trust Repair phase. We included a larger number of AI-correct tasks in this phase to have a longer runway to observe when and how trust levels from participants returned to their previous

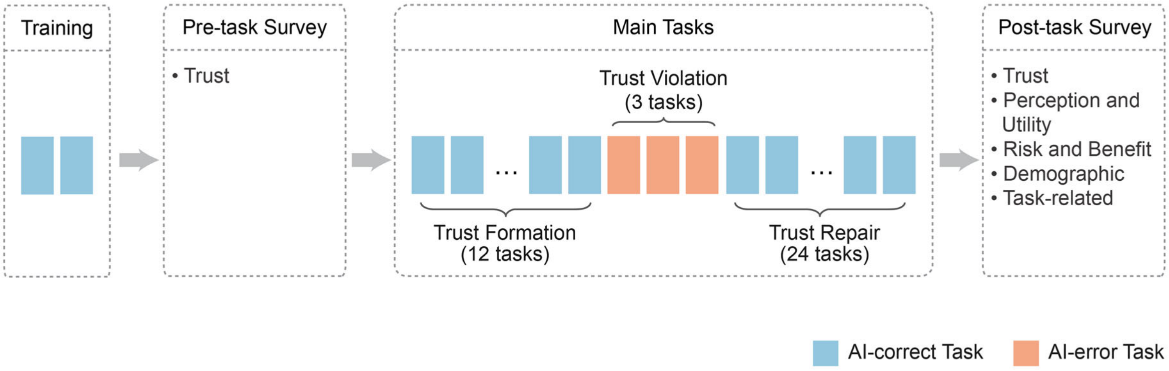


Figure 2. Experiment procedure of the study. During the main tasks, human trust in AI was manipulated into three trust phases (Trust Formation, Trust Violation, and Trust Repair) by arranging the sequence of AI-correct and AI-error tasks.

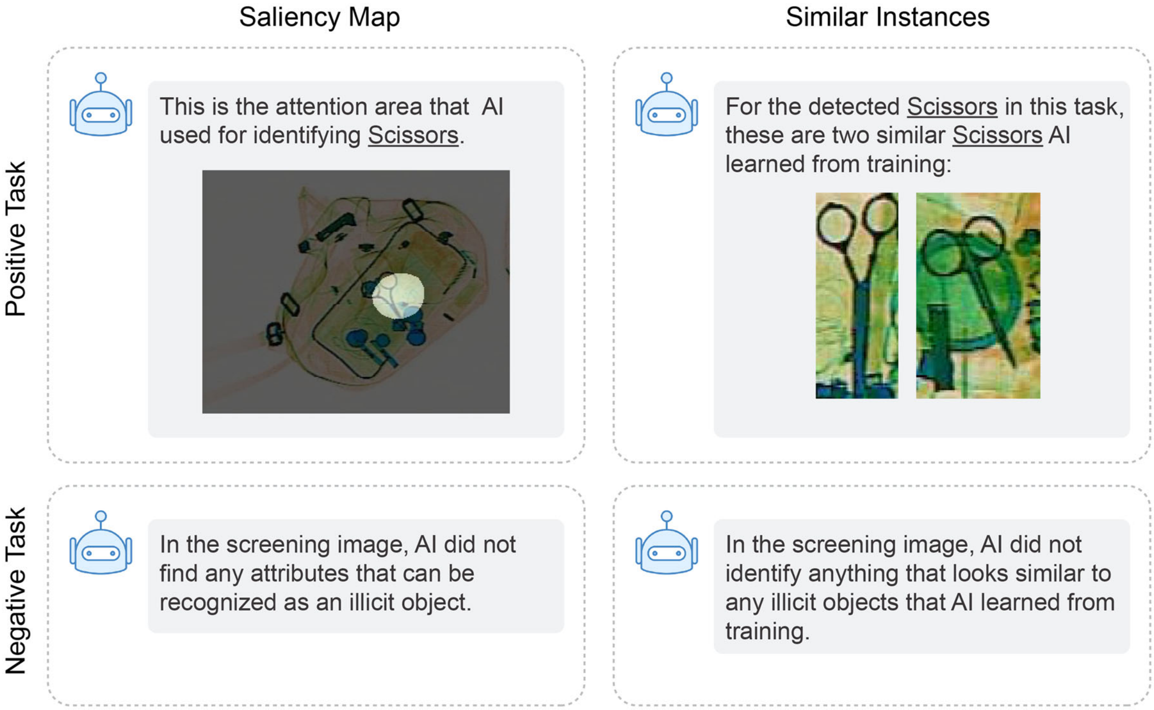
state. Both Trust Formation and Trust Repair phases consisted of easy and difficult tasks to emulate a practical but engaging task environment.

3.1.4. Experimental condition

This study also explored the effect of three factors on the evolution of human-AI trust over time using a between-subjects design: XAI representation, type of AI error, and strategy of conveying XAI.

3.1.4.1. XAI representation. We focused on the XAI techniques that interpret AI results through localization methods, providing explanations for each AI output. Since both feature-based and example-based explanations have been shown to effectively calibrate human-AI trust (Wang & Yin, 2022; Yang et al., 2020), our work included the Saliency Map to represent feature-based explanations, utilizing the class activation mapping (CAM) algorithm (Zhou et al., 2016), and the Similar Instances, generated by the DeepImageSearch (Verma, 2021) algorithm, to represent example-based explanations. We also incorporated a control condition (No XAI), where no explanation was provided to participants, to benchmark the effects of two XAI visualizations.

As shown in Figure 3, for positive baggage screening tasks where AI detected an illicit object, participants under the Saliency Map condition were presented with explanations by viewing the shadowed baggage screening task image, with the most critical areas contributing to AI detection highlighted. The saliency map was also accompanied by a text description: "This is the attention area that AI used for identifying [particular illicit object detected]." Under the Similar Instances condition, for each AI-detected illicit object in the current task, participants were provided with two of the most similar illicit objects from the same class that AI learned from training. Participants were also informed that "For the detected [particular illicit object detected] in this task, these are two similar [particular illicit object detected] AI learned from training. The usability of Saliency Map and Similar Instances explanations for negative detection results has not been well studied through human subject studies. Such explanations may pose challenges in helping people understand why the AI did not detect any illicit object. For example, in the Saliency Map condition, an empty area in the baggage would be highlighted as the attention area that aids AI decision-making. Whereas the Similar Instances may involve two pieces of baggage containing similar benign items. These explanations could be demanding for participants to understand why no illicit object was detected. Thus, in our study, participants did not receive visual explanations for negative baggage screening results. Instead, they were provided with text explanations. Under the Saliency Map condition, participants were notified: "In the screening image, AI did not find any attributes that can be recognized as an illicit object." Under the Similar Instances condition, participants were told: "In the screening image, AI did not identify anything that looks similar to any illicit objects that the AI learned from training."



This is the attention area that AI used for identifying Scissors.

For the detected Scissors in this task, these are two similar Scissors AI learned from training:

In the screening image, AI did not find any attributes that can be recognized as an illicit object.

In the screening image, AI did not identify anything that looks similar to any illicit objects that AI learned from training.

Figure 3. Two representations of XAI (Saliency Map, Similar Instances) were applied to positive (AI detected an illicit object) and negative (AI detected no illicit object) baggage screening tasks.

3.1.4.2. Type of AI error. Since different types of AI errors and decision severity can affect human trust in AI to varying degrees (Baughan et al., 2023; Tolmeijer et al., 2021; Wang et al., 2022), this study investigated the impact of two common types of AI errors that occur in real-world baggage screening tasks: False Alarm and Miss Signals. For brevity, we will refer to “Miss signals” as “Miss” in subsequent discussions. False Alarm refers to the AI mistakenly identifying any benign objects as illicit, which may result in unnecessary screening and time delays. Miss denotes that the AI fails to detect the presence of an illicit object, which could pose a risk to public safety, yielding more severe consequences than False Alarm. We incorporated both error types in the study to examine their impacts on the Trust Violation phase and the subsequent Trust Repair phase. Each participant was presented with three successive AI errors of the same type in illicit object detection (Supplementary Appendix A).

Our study excluded other screening errors such as misclassification (e.g., identifying a knife as a wrench) or partial detections (e.g., detecting one illicit item but missing another), which could introduce additional variables and complicate the study design. Focusing on the two most typical screening errors can help reduce people’s cognitive overload and enable people to form clear mental models of the AI’s performance.

3.1.4.3. Strategy of conveying XAI. Drawing insights from social science research, explanations involve a social process of conveying knowledge through communication. Explanations are not simply transparency displays, presentations of a machine’s reasoning, or providing information unilaterally. Rather, a responsive explanation provides contrastive information that is relevant to the question in mind (Miller, 2019). To address these social considerations of what an explanation should be and the context surrounding what “makes” an explanation, our study investigated three distinct strategies for conveying XAI visualizations to people: Responsive-all (RA), Non-responsive-all (NRA), and Non-responsive-difficult (NRD). In each baggage screening task under XAI conditions, participants in the NRA group were informed of the AI detection result and the associated explanation directly. Yet, participants in the RA group received XAI only when they explicitly requested it, indicating they had a question in mind about the AI’s recommendation. The NRD condition was introduced based on observations from our pilot study (Section 3.4), where participants tended to request XAI more frequently during difficult tasks. In the NRD group for the formal study, XAI visualizations were

Table 1. Fourteen (14) experimental conditions examined in the study.

XAI representation	Type of AI error	Strategy of conveying XAI
Saliency Map	False Alarm	Responsive-all (RA)
Saliency Map	False Alarm	Non-responsive-all (NRA)
Saliency Map	False Alarm	Non-responsive-difficult (NRD)
Saliency Map	Miss	Responsive-all (RA)
Saliency Map	Miss	Non-responsive-all (NRA)
Saliency Map	Miss	Non-responsive-difficult (NRD)
Similar Instances	False Alarm	Responsive-all (RA)
Similar Instances	False Alarm	Non-responsive-all (NRA)
Similar Instances	False Alarm	Non-responsive-difficult (NRD)
Similar Instances	Miss	Responsive-all (RA)
Similar Instances	Miss	Non-responsive-all (NRA)
Similar Instances	Miss	Non-responsive-difficult (NRD)
No XAI	False Alarm	Non-responsive-all (NRA)
No XAI	Miss	Non-responsive-all (NRA)

automatically provided only during difficult baggage screening tasks, whereas no explanation was made available during the easy tasks.

To summarize, our study employed a mixed factorial design with three between-subject factors and one within-subject factor. The between-suject factors were XAI representation (three levels), type of AI error (two levels), and strategy of conveying XAI (three levels). A total of 14 experimental conditions were examined (Table 1): of these, there were $2 \times 2 \times 3 = 12$ conditions presenting XAI information, across two XAI representations (Saliency Map and Similar Instances), two types of AI errors (False Alarm and Miss), and three XAI delivery strategies (RA, NRA, and NRD); two No XAI groups under the NRA condition for each type of AI error were included as control conditions ($1 \times 1 \times 2 = 2$), to benchmark the effects of XAI representations. In addition, a within-subject factor, trust phase (three levels), was included to enable repeated measurements across time for each participant.

3.2. Participants

Participants in the study were recruited through Prolific and randomly assigned to one of the experimental conditions. Following recommendations from prior work showing that factorial designs can achieve strong statistical power with relatively few participants per condition (Collins et al., 2014), we adopted a balanced design—recruiting 30 participants per condition in the pilot study and approximately 20 per condition across 14 conditions in the formal study (Table B.1). A total of 60 U.S. residents (Age: $\mu = 36.5$, $\sigma = 12.6$ years) completed the pilot study, including 37 females and 23 males. The formal study included 281 U.S. residents (Age: $\mu = 37.3$, $\sigma = 11.1$ years), comprising 121 females, 158 males, 1 non-binary, and 1 who selected "other." Those who selected "other" were asked to provide a self-reported description. Participants in the formal study were compensated with a base payment of $8 for completing the study, with an average completion time being approximately 29 min. They were also incentivized with a bonus of $0.15 for a correct detection in each baggage screening task. One participant's data was excluded from the analysis for the formal study because the participant agreed with the AI on all three AI-error tasks, indicating a severe lack of understanding or attention. This resulted in a final sample of 280 participants included in the analysis.

3.3. Procedure and task

3.3.1. Study procedure

As outlined in Figure 2, participants were first asked to complete the study information sheet, where they were informed of the bonus structure and received instructions for the baggage screening tasks. After that, a tutorial video and following attention-check quiz were provided to ensure they understood both the task requirements and the incentive rules. Also, participants were informed of AI performance during the tutorial (accuracy: 84.7%, various error rates). Only participants who passed the quiz could proceed to the task training, where they practiced illicit object detection in one easy and one difficult AI-correct baggage screening task without affecting the bonus. This practice session familiarized participants with the task interface and procedures. After that, a pre-task trust survey (Jian et al., 2000) was administered to measure participants' initial trust in AI, and then participants proceeded to the main baggage screening tasks.

As described in Section 3.1, the main baggage screening tasks were managed as follows: 12 AI-correct baggage screening tasks were presented at first, serving as Trust Formation. Then, three consecutive salient AI errors were introduced, leading to Trust Violation. Finally, an additional 24 AI-correct tasks were displayed for Trust Repair. To simulate a time-constrained environment, participants were informed of a 30-minute time limit during the tutorial and at the beginning of the main tasks. A countdown clock was also displayed throughout the main tasks to maintain this pressure. After performing 39 main baggage screening tasks, participants were asked to fill out post-task surveys to assess their trust in AI (Jian et al., 2000), perceived understanding and utility (Cheng et al., 2019), risk and benefit perception (Weber et al., 2002), and demographic information. Following the recommendation in prior work (Gutzwiller et al., 2019), the response items in both pre-task and post-task trust surveys were randomized to eliminate maturation effects. Furthermore, participants needed to revisit the same three AI-errors tasks and determine whether they considered the AI result to be correct or not. This result helped us further confirm the occurrence of trust degradation in the Trust Violation phase (Table A.1). In the end, a study result was concluded for each participant with a report detailing decision accuracy for each baggage screening task and the corresponding bonus.

3.3.2. Task prototype

Our study adopted a chatbot and gathered data from participants through a conversational survey which has been shown to generate higher quality responses than traditional form-based surveys (Xiao et al., 2020). As shown in Figure 1, each baggage screening task began by presenting participants with a baggage screening image along with the AI detection result. Depending on the participant's assigned experimental conditions, participants either received or did not receive XAI for the AI result. Participants in the No XAI group were directly asked to indicate their agreement with the AI detection result by responding to the question, "Do you agree with the AI detection?" In contrast, participants in the XAI groups (Saliency Map, Similar Instances) received explanations of the AI detection through one of the three different strategies (RA, NRA, NRD) before being asked for their agreement with the AI result. In particular, under the RA condition, the AI initiated the conversation by asking, "Do you want to know why AI made the detection result?" Only participants who requested XAI were offered the explanation information. However, XAI was automatically shown to participants at the beginning of the chat for all tasks under the NRA condition, but difficult tasks only under the NRD condition. After receiving the XAI, participants needed to determine if they agreed with the AI detection result or not, and then make the final decision by answering the question, "Do you want to open the baggage and confirm all illicit objects?" Participants could choose to either open the baggage and submit the ground truth, accept the model detection if they agreed with it beforehand, or make the detection on their own if they disagreed with the model detection beforehand.

3.4. Pilot study

To validate the study design and task selection, participants were asked to perform baggage screening on either 25 tasks (including 24 AI-correct tasks and one AI-error task) or 27 tasks (comprising 24 AI-correct tasks and three AI-error tasks). These tasks were presented in a fixed sequence of 12 AI-correct tasks, followed by three or one AI-error task, and then another 12 AI-correct tasks. We conducted two AI-assisted experiments under the RA condition for each type of AI error, using the Similar Instances XAI. Additionally, we implemented a control experiment where participants performed 27 baggage screening tasks without any AI assistance. We followed predefined criteria to distinguish between difficult and easy tasks, manually selected 54 tasks covering both positive and negative cases, and randomly sampled them for each group in the pilot study. After confirming their difficulty levels based on human performance without AI assistance, we selected 36 of these tasks for use in the formal study. The pilot study results (Supplementary Appendix C) validated task difficulty and showed that presenting three successive AI-error tasks led to lower agreement with the AI compared to a single error, although agreement levels returned to baseline afterward in both conditions. This finding supports our decision to use three consecutive AI errors to induce trust violations in the formal study. We also found that participants were more likely to request XAI when encountering difficult tasks. This observation inspired a new XAI delivery strategy for the formal study, i.e., NRD. Pilot study data were analyzed using JMP®, with more detailed results included in the supplemental materials.

3.5. Measurement

To make comprehensive comparisons of human trust in AI and human-AI collaborative performance over time across different experimental conditions, we employed the following measurements:

3.5.1. Behavioral trust

There are many potential measures of "behavioral trust," some that have been widely used across studies and others that are more bespoke to the task environment (Kohn et al., 2021). Ultimately, these proxy measures of trust must be interpreted in context. In this work, we operationalize "behavioral trust" as the willingness to act based on the information provided by the AI (Chen & Sundar, 2023). Participants performed baggage screening tasks with AI assistance. They could choose to agree with the AI detection results, make independent decisions, or spend additional time opening the baggage to verify the ground truth. Under the RA condition, participants also had the option to request or deny explanations for the AI results. Therefore, human behavioral trust was evaluated by observing participants' behaviors and interactions with the system during the decision-making process, focusing on their compliance with AI, but also their confidence in their decisions, and their interest in seeking explanations for AI outcomes.

- **Compliance with AI (*Agreement*).** Human compliance with AI describes people directly using AI results when making final decisions. This differs from reliance behaviors, which are more passive, like doing nothing even when the AI provides a false negative result (Meyer et al., 2014). In general, agreement with AI has been commonly used as a key metric for measuring human reliance and compliance with AI systems (Yin et al., 2019; Zhang et al., 2020), with higher levels of compliance or reliance being associated with higher levels of trust (Lee & See, 2004). Thus, our work also applied this measure to assess humans' compliance with the AI for each task.
- **Confidence in Decision (*Open Baggage*).** Higher self-confidence in a task tends to reduce trust in automation (Lee & Moray, 1994), while greater confidence in AI decisions increases trust in the AI (Chong et al., 2022). In this work, rather than asking for self-reported confidence ratings for each task which could become tedious rather quickly, we instead assessed participants' confidence level as a binary measure, based on their behaviors of opting to open the baggage or not. If participants chose to spend extra time opening and inspecting the baggage, it indicated a lack of confidence in either the AI's detection result or their own detection results (a desire for more information to further verify). Conversely, opting not to open the baggage suggested confidence in either the AI or their own detection results.
- **Interest in Seeking Explanation for AI Outcome (*See XAI*).** Under the RA condition, opting for an explanation serves as a measure of behavioral trust, similar to risk-taking in a relationship (Mayer et al., 1995; Simpson, 2007), but with dual interpretations. When interpreted in context, this behavioral measure may provide further insights on a person's trust state (Kohn et al., 2021). Specifically, participants may not feel the need to request additional explanations if they have already developed strong trust or distrust. Therefore, choosing to see XAI during a phase of the high level of human trust (e.g., the Trust Formation or Trust Repair phase) could imply a decline in human-AI trust because they are engaging in preliminary verification behaviors short of opening the baggage for more concrete (and costly) verification. Conversely, requesting XAI during a low level of trust phase (e.g., the Trust Violation phase) could indicate a potential rebound in trust, as participants actively seek more information to better understand the AI decisions.

3.5.2. Perceived trust

Participants were asked to rate their trust-related perceptions through a trust questionnaire (Jian et al., 2000) before (pre-task) and after (post-task) completing the baggage screening tasks. The questionnaire identified 12 distinct factors that characterize how people perceive trust in an automated system, comprising both positively worded (statements that express a favorable opinion about AI) and negatively worded (statements that express an unfavorable opinion about AI) questions, and each item response is on a 7-point Likert scale. The self-reported responses from the trust questionnaire were used as subjective measures of perceived trust.

3.5.3. Decision accuracy

For each task, accuracy was defined as the final detection result being correct only if it identified all actual illicit objects without including any nonexistent ones. Otherwise, the result is regarded as inaccurate. Decision accuracy serves as the primary metric to evaluate human-AI collaborative performance.

3.6. Statistical analysis

Separate analyses were performed for the following participant groups (see Table B.1 for detailed group sizes across the 14 experimental conditions):

1. NRA groups ($N = 121$): Participants in the six NRA groups were analyzed to examine whether the availability of XAI, i.e., No XAI vs. With XAI (Saliency Map, Similar Instances), influenced human-AI trust and collaborative performance;
2. XAI groups ($N = 232$): Among the 240 participants assigned to the 12 XAI conditions (including Saliency Map and Similar Instances), eight participants were excluded because they did not opt to view XAI explanations both before and after AI errors. The remaining 232 participants were included to explore the effects of different XAI representations, types of AI error, and strategies of presenting XAI on human-AI trust and collaborative performance;
3. RA groups ($N = 71$): 79 participants in four RA groups were shown XAI only upon request. 8 participants were excluded because they did not opt to view XAI explanations both before and after AI errors. This analysis focused on how XAI representation and type of AI error impact participants' willingness to request explanations (i.e., *See XAI*).

We used Generalized Linear Mixed Models (GLMM) with two-way interaction terms between the three trust phases and experimental factors (XAI representation, type of AI error, strategy of conveying XAI) to compare behavioral trust and performance over time across experimental conditions. This approach maintained a consistent statistical modeling strategy for the different response variables. A crude version of the GLMM model was defined as:

$$g(\mathbb{E}[y|u]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + Zu \tag{1}$$

where $g(.)$ is a transformation function for the expected value of response y, u refers to a random effect of different participants, β_i denotes the intercept or fixed effects, and Z represents the random effects design matrix. When y refers to *Agreement*, *Open Baggage*, *See XAI*, *Decision Accuracy*, the transformation function is a binomial distribution. X_1 denotes a set of experimental factors that could contain the type of XAI representation, type of AI error, and type of strategy of presenting XAI based on different datasets. X_2 denotes trust phases. $(X_1 \cdot X_2)$ are interactions between the set of experimental factors and trust phases. In addition, we used GLMM models with a similar model specification to analyze participants' perceived trust through their responses to the trust survey. Given that trust in automation is a multi-faceted construct (Chiou & Lee, 2023; Lee & See, 2004), and our interest lies in how specific components of trust and distrust are influenced by experimental factors over time, we analyzed each survey item individually. When modeling participants' perceived trust, the transformation function $g(.)$ was adjusted to be the cumulative multinomial distribution. Although X_1 denotes the same experimental factors as mentioned above, X_2 is a binary variable to indicate pre- or post-task survey and $(X_1 \cdot X_2)$ are interactions between the set of experimental factors and the time of requesting survey.

4. Results

In this section, we present the results of comparing participants' behavioral trust indicators, trust perceptions, and decision performance across different experimental conditions. We used SAS® to estimate the GLMMs and to perform tests of significance on the effects of interest. For statistically significant results, we report both p-values and odds ratios; for non-significant effects, only p-values are provided due to minimal differences between groups. All reported data are included in Supplementary Appendix D, and detailed statistical results can be assessed in OSF.[3]

4.1. Behavioral trust

4.1.1. Agreement

In the NRA group (Tables D.4 – D.6), we found no significant difference in the frequency of agreement with AI results between No XAI and Saliency Map conditions ($p = 0.496$). However, both frequencies were significantly lower than that observed under the Similar Instances condition (No XAI vs. Similar Instances $p = 0.004, OR = 0.592$ $[0.414, 0.846]$; Saliency Map vs. Similar Instances $p = 0.043, OR = 0.681$ $[0.470, 0.988]$). This suggests that observing Similar Instances in each baggage screening task results in higher human compliance with AI in comparison to Saliency Map or No XAI. Participants showed the lowest levels of agreement with AI results during the Trust Violation phase, compared to Trust Formation ($p < 0.0001, OR = 72.683$ $[47.604, 110.976]$) and Trust Repair ($p < 0.0001, OR = 86.559$ $[57.356, 130.631]$). Statistical tests of significance revealed no significant interaction effect between XAI representation and the trust phase on human agreement with AI results ($p = 0.203$).

For participants exposed to XAI in the Trust Formation and Trust Repair phases, *Agreement* levels were similar between these phases (Figure 4). However, during the Trust Violation phase, *Agreement* decreased more with the Saliency Map than with the Similar Instances ($p = 0.0001, OR = 0.300$ $[0.163, 0.555]$). This pattern indicates an interaction effect between the XAI representation and the trust phase on human compliance (i.e., *Agreement*) with AI, suggesting that the Saliency Map results in more careful decisions when AI errors occur. Additionally, the type of AI errors and the trust phase were found to jointly impact participants' agreement with AI outputs (Figure 4, middle): During the Trust Violation phase, participants agreed with AI detection results at similar rates for both AI error types. However, after observing AI errors (i.e., Trust Repair), participants who encountered Miss errors showed a higher level of agreement with AI outcomes than those who encountered False Alarm errors ($p = 0.045, OR = 1.209$ $[1.004, 1.458]$). Since visual explanations were provided only for False Alarm AI errors, this finding suggests that explanations for AI predictions could reduce human compliance with AI after participants observe incorrect results. However, the strategy of presenting XAI did not significantly affect participants' agreement with the AI ($p = 0.928$).

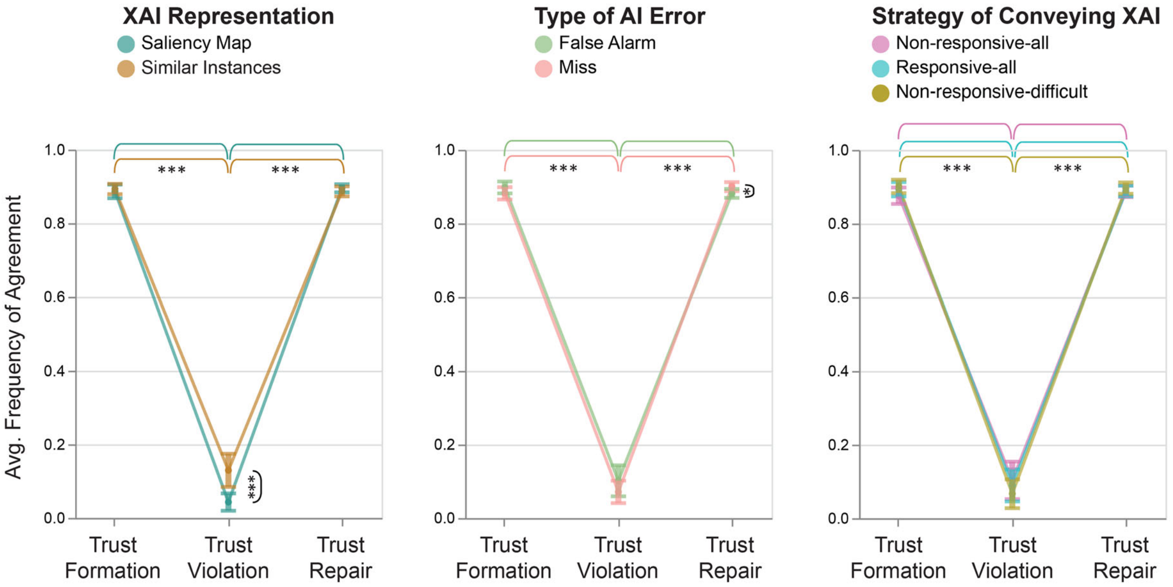


Figure 4. Effects of different XAI representations, types of AI error, and strategies of conveying XAI on participants' average frequency of agreeing with AI results across three trust phases. This figure covered participants who viewed XAI both before and after AI errors. Dots indicate average frequencies and error bars indicate 95% confidence intervals, and detailed statistics are provided in Supplementary Tables D.1–D.3 ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

4.1.2. Open baggage

For participants in the NRA groups (Tables D.10, D.11), frequencies of *Open Baggage* were comparable among Similar Instances, Saliency Map, and No XAI conditions, suggesting that explaining AI results has no impact on human confidence in AI decisions ($p = 0.302$).

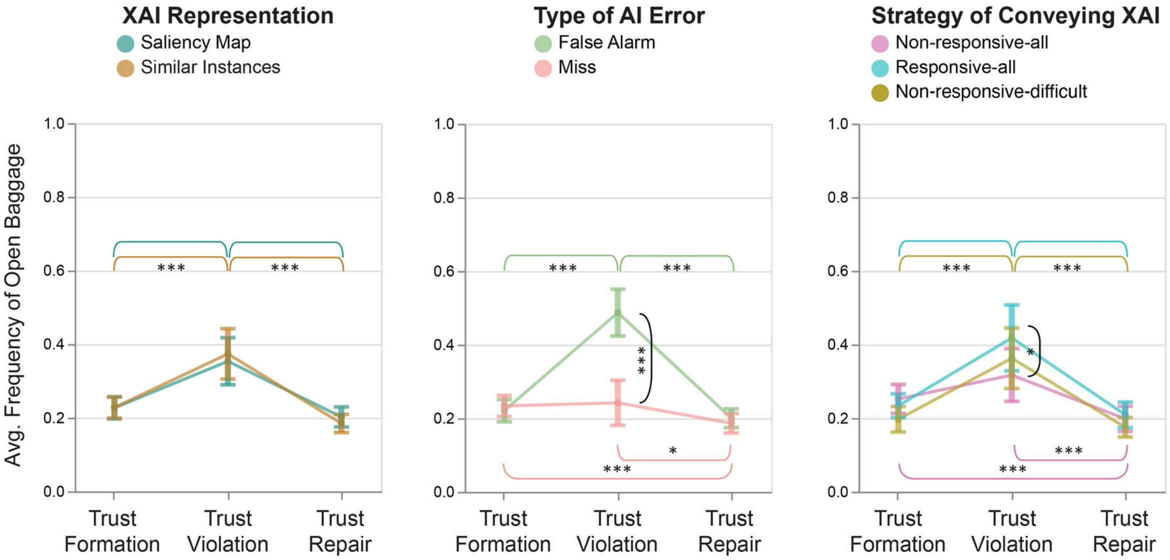


Figure 5. Effects of different XAI representations, types of AI error, and strategies of conveying XAI on participants' average frequency of requesting open baggage across three trust phases. This figure covered participants who viewed XAI both before and after AI errors. Dots indicate average frequencies and error bars indicate 95% confidence intervals, and detailed statistics are provided in Supplementary Tables D.7–D.9 (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

As shown in Figure 5 (left), statistical comparisons of the *Open Baggage* action among participants who viewed XAI before and after AI errors further confirm that there are no significant differences between the two XAI representations ($p = 0.922$). Differences between the two types of AI errors (Figure 5, middle) were significant and interacted with the trust phase: During Trust Violation, participants chose *Open Baggage* more frequently when observing False Alarm AI errors than Miss AI errors ($p < 0.0001, OR = 3.438\ [2.309, 5.120]$). However, the frequency of opening baggage realigned in the Trust Repair phase for both AI error types. This observation suggests that participants encountering False Alarm AI errors were less confident in their decisions, as they were provided specific explanations for the AI's incorrect results. Thus, XAI could help calibrate human confidence and encourage more cautious decision-making in the presence of AI errors. Additionally, the strategy of conveying XAI was observed to affect participants' choice of *Open Baggage* in some trust phases. During the Trust Formation phase, participants who received XAI for each task under the NRA condition were more likely to request opening baggage than those under the NRD condition ($p = 0.0415, OR = 1.411\ [1.013, 1.965]$). However, the frequency of opening baggage in these two conditions again realigned when AI errors occurred. Conversely, participants in the RA group were more inclined to open baggage than those in the NRA group when encountering AI errors ($p = 0.0266, OR = 1.739\ [1.067, 2.833]$). Despite this, no significant difference in the frequency of opening baggage was found between participants under the RA condition and those under the NRA and NRD conditions, both before (RA vs. NRA $p = 0.602$; RA vs. NRD $p = 0.145$) and after (RA vs. NRA $p = 0.489$; RA vs. NRD $p = 0.098$) experiencing AI mistakes.

4.1.3. See XAI

To assess the participants' tendencies in seeking explanations for AI outputs, we provided participants in the RA group with the option to choose whether or not to see the XAI. Participants exposed to both Similar Instances and Saliency Map showed similar patterns in their choices. Although the overall frequency of choosing to see Similar Instances was slightly higher for Saliency Map, this difference was not statistically significant ($p = 0.538$). However, statistical analysis revealed a significant interaction effect between the type of AI error and the trust phase. As shown in Figure 6, the desire to seek explanations for AI outputs decreased when participants encountered Miss AI errors but increased under the False Alarm condition. Participants were also more likely to request XAI when encountering False Alarm errors than Miss errors ($p < 0.0001, OR = 30.318\ [11.778, 78.039]$). This pattern could be because AI provided only textual explanations for Miss errors, stating no illicit object was detected – information already familiar to participants due to prior experience. After AI errors occurred, the frequency of requesting XAI was aligned between the two groups. Additionally, the proportion of participants selecting *See XAI* in the Trust Repair phase was lower than in the Trust Formation phase ($p < 0.0001, OR = 0.428\ [0.348, 0.527]$), suggesting that observing AI mistakes in visually easier tasks reduces participants' proclivity to understand the reasons behind AI decisions, potentially signaling a decline in human-AI trust.

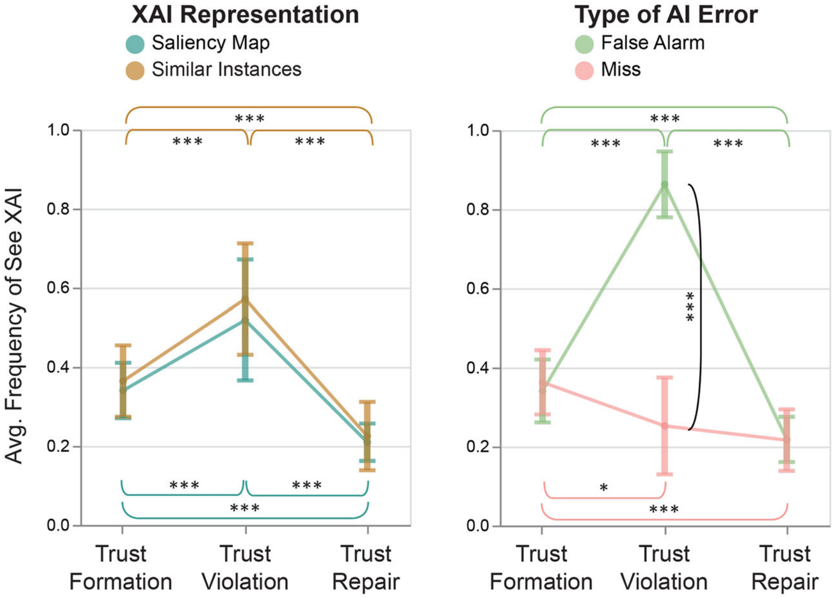


Figure 6. Effects of different XAI representations and types of AI error on participants' average frequency of requesting to see XAI across three trust phases. This figure includes participants in the RA group. Dots indicate average frequencies and error bars indicate 95% confidence intervals, and detailed statistics are provided in Supplementary Tables D.12 and D.13 (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

4.2. Perceived trust

For each trust response, we explored the main effects and the 2-way interactions between the time of trust perception measurement (pre-task vs. post-task) and experimental factors, including types of XAI representation, type of AI error, and strategy of conveying XAI. We analyzed the trust perception of 232 participants who viewed XAI before and after AI errors.

Most trust survey responses showed significant differences between the two repeated measures (pre-task vs. post-task), except for the statement "*The AI behaves in an underhanded manner*" ($p = 0.319$). For the items with significant differences, positively worded items in the post-task survey were rated higher than in the pre-task survey, while negatively worded items were rated lower. This suggests that, overall, participants' perceived trust in AI remains consistent and even increased after completing 39 baggage screening tasks, despite observing AI errors.

Further analysis of the experimental conditions' effects on perceived trust revealed that XAI representations significantly influenced perceived familiarity with the AI (as a response to the survey item "*I am familiar with the AI*"). As shown in Figure 7, participants exhibited consistent familiarity levels with AI before the baggage screening tasks ($p = 0.451$). However, after completing 39 tasks, AI was perceived as more familiar by participants exposed to the Saliency Map than those exposed to Similar Instances ($p = 0.011, OR = 2.895\ [1.277, 6.565]$). While the 121 participants in the NRA group did not show significant differences in perceived trust when comparing the No XAI condition with the two XAI conditions ($p > 0.05$ for all survey items).

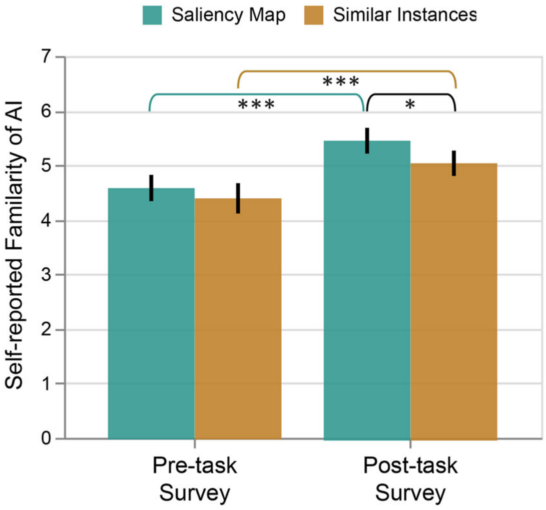


Figure 7. The effect of different XAI representations on the average self-reported score of familiarity item (*"I am familiar with the AI"*) in the pre-task and post-task trust survey. There is an active interactive effect between the XAI representation and the time of measuring familiarity. This figure covered participants who viewed XAI both before and after AI errors. Bars indicate average self-reported score and error bars indicate 95% confidence intervals, and detailed statistics are provided in Supplementary Tables D.14 and D.15 ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Among participants exposed to XAI before and after AI errors, those who were exposed to false misses perceived the AI as having more integrity ($p = 0.007, OR = 4.533$ $[1.508, 13.631]$) and being less harmful ($p = 0.041, OR = 0.431$ $[0.192, 0.967]$), in comparison to those who were exposed to false alarms (Figure 8). This is despite the fact that Miss AI errors result in more severe consequences compared to False Alarm errors. One possible explanation is that providing detailed and concrete explanations for AI outcomes during False Alarm errors may inadvertently emphasize the AI's limitations, prompting people to question the

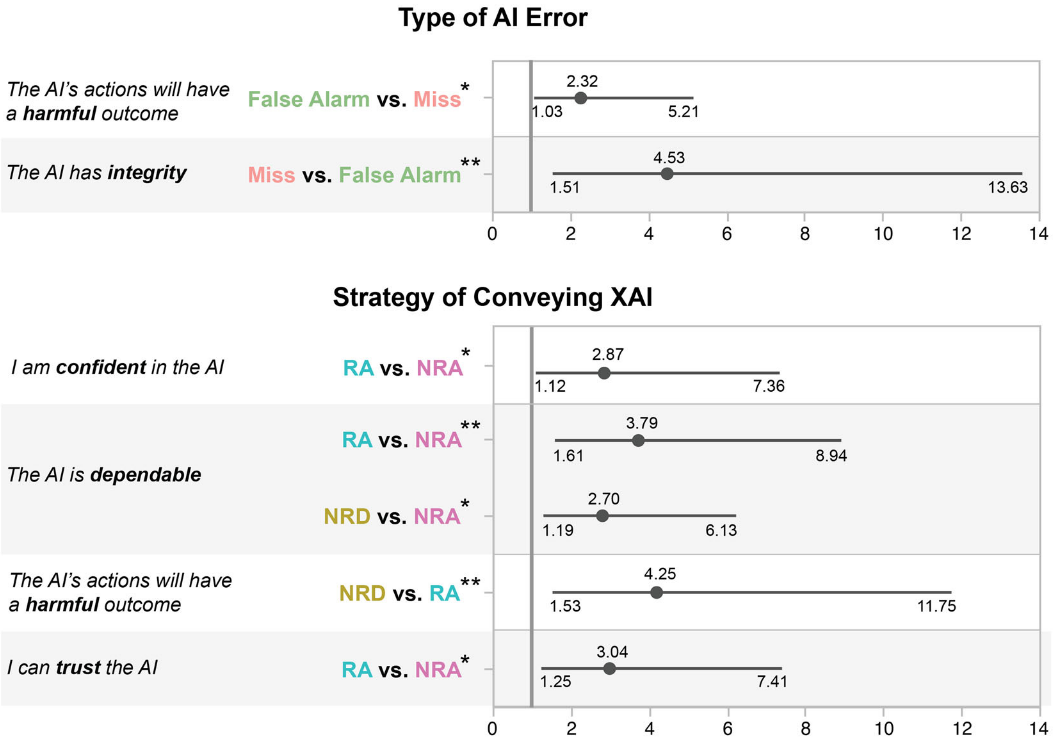


Figure 8. A sample of self-reported ratings of trust survey items from participants who viewed XAI both before and after AI errors. The 95% CI interval plots show the odds ratios between the two types of AI error and three strategies of conveying XAI. Statistical significant differences are included exclusively, and detailed statistics are provided in Supplementary Tables D.16–D.25 ($^*p < 0.05$, $^{**}p < 0.01$).

reasoning behind its decisions. As a result, increased awareness of these errors could undermine human trust in the AI system (Bliss & Fallon, 2006). Participants who encountered Miss errors also reported slightly more familiarity with the AI than those who experienced False Alarm errors ($p = 0.085$). Further investigation of the data revealed that encountering False Alarm errors led to greater wariness of AI than Miss errors after the tasks ($p = 0.052$), despite similar initial wariness levels.

Different strategies for communicating XAI also significantly affected trust perception (Figure 8). The RA group inspired more confidence ($p = 0.029, OR = 2.869$ $[1.118, 7.364]$) and trust ($p = 0.015, OR = 3.042$ $[1.248, 7.413]$) in AI than the NRA group. AI in both RA and NRD groups was regarded as more dependable than in the NRA group (RA vs. NRA: $p = 0.003, OR = 3.794$ $[1.609, 8.946]$; NRD vs. NRA: $p = 0.017, OR = 2.703$ $[1.193, 6.126]$). Participants in the RA group perceived AI actions as less harmful compared to the NRD group ($p = 0.006, OR = 0.236$ $[0.0851, 0.652]$. This suggests that providing XAI when needed is more effective in enhancing overall perceived trust in the AI.

4.3. Decision accuracy

To compare human-AI collaborative performance between conditions with and without XAI representations, we examined the decision accuracy of 121 participants in the NRA group (Tables D.28, D.29). We observed that decision accuracy was nearly identical for participants in the No XAI condition and those in the XAI conditions (No XAI vs. Saliency Map $p = 0.090$; No XAI vs. Similar Instances $p = 0.404$; Saliency Map vs. Similar Instances $p = 0.326$).

Figure 9 shows a significant interaction effect between the type of XAI representation and the trust phase among participants who viewed XAI before and after AI errors. In the Trust Formation phase, decision accuracy was similar for both two XAI conditions ($p = 0.980$). However, during the Trust Violation phase, decision accuracy significantly declined with the Similar Instances condition compared to the Saliency Map condition ($p = 0.001, OR = 0.255$ $[0.112, 0.578]$). After experiencing AI errors, decision accuracy in both XAI conditions improved during the Trust Repair phase, with the Similar Instances showing slightly better performance than Saliency Map, though this difference was not statistically significant ($p = 0.137$). The consistent performance across the XAI conditions during the Trust Formation and Trust Repair phases may be attributed to the fact that the AI was completely accurate during these phases and the tasks were relatively easy to verify manually. Additionally, neither the type of AI error ($p = 0.237$) nor the strategy of conveying XAI ($p = 0.306$) was found to significantly affect decision accuracy in this study.

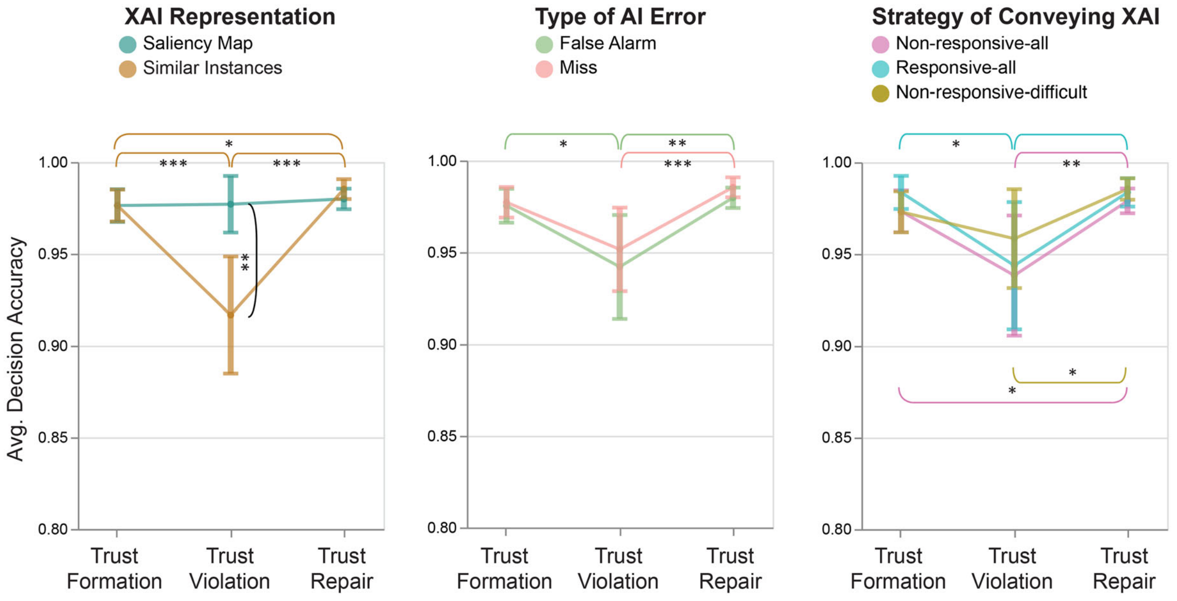


Figure 9. Effects of different XAI representations, types of AI error, and strategies of conveying XAI on participants' average decision accuracy across three trust phases. This figure covered participants who viewed XAI both before and after AI errors. Dots indicate the average frequencies and error bars indicate 95% confidence intervals, and detailed statistics are provided in Supplementary Tables D.26 and D.27 (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

4.4. Summary of key results

4.4.1. XAI representation

- Consistently provided Similar Instances explanation promoted greater human compliance with AI (i.e., more agreement with AI) than Saliency Map or No XAI.
- Saliency Map led to more appropriate human compliance with AI during AI errors (i.e., less agreement with AI) compared to Similar Instances.
- The decision accuracy of Similar Instances was comparable to Saliency Map during the Trust Formation and Trust Repair phases but significantly lower than that of Saliency Map during the Trust Violation phase.
- Saliency Map encouraged a greater sense of familiarity with AI among participants after completing 39 baggage screening tasks compared to those exposed to Similar Instances, despite similar familiarity levels before the task.

4.4.2. Type of AI error

- After observing AI errors, participants showed higher agreement with AI outcomes for Miss errors compared to False Alarm errors, suggesting that more concrete explanations for AI results may decrease human compliance with AI after errors are noticed.
- Participants opened baggage more frequently when encountering False Alarm errors than Miss errors, though this behavior balanced out during Trust Repair. It implied that concrete explanations of AI outcomes encourage more cautious and calibrated decision-making.
- Participants were more tended to request explanations (XAI) when encountering False Alarm compared to Miss errors, likely due to the extra explanatory information being only provided for detected objects.
- Experiencing Miss errors led participants to perceive AI as having more integrity and being less harmful than False Alarm errors.

4.4.3. Strategy of conveying XAI

- The RA approach is better rated as promoting trust-related perceptions of AI in the measures of “confident,” “dependable,” and “trust” compared with the NRA approach. Additionally, participants with the NRD approach found AI to be more dependable than with the NRA approach, while more harmful than with the RA approach. This indicates that providing XAI “when needed” was more effective in improving perceived AI trustworthiness.

5. Discussion

We investigated how human trust in AI evolves in a synthetic baggage screening environment where participants, acting as security screeners, worked alongside an AI counterpart to jointly detect prohibited objects in 2D X-ray images. Our objective was to understand how different representations of explanations for AI predictions (i.e., XAI), strategies of communicating these explanations, and two types of AI error affect overall trust evolution during human-AI collaboration. In this work, we designed AI output sequences based on the premise that human trust in AI follows a progression influenced by AI performance. By carefully designing the sequence of tasks, we aimed to establish an initial foundation of trust in the AI through continuous AI-correct tasks (Trust Formation), intentionally disrupting it through deliberate consecutively AI errors that are obvious to people (Trust Violation), and subsequently attempt to rebuild trust by another set of AI-correct tasks (Trust Repair). This sequential manipulation enabled us to gain nuanced insights into how the types of XAI representations, XAI delivery strategies, and types of AI errors affect human trust level and human-AI collaboration across different phases of human-AI interaction. In this section, we summarize our findings, discuss their implications, outline the limitations of our study, and suggest directions for future research.

5.1. XAI in the presence of AI error

5.1.1. RQ1: Effect of explanation Type on trust during AI errors

In the Non-responsive-all (NRA) group, participants who were continuously shown example-based explanations (Similar Instances) exhibited a higher rate of *Agreement* with the AI's decisions than those who saw feature-based explanations (Saliency Maps) or no explanations. This trend suggests that frequent exposure to concrete, past, and even similar examples, can more rapidly build initial behavioral trust by leveraging the human tendency for analogical reasoning i.e., people are more likely to intuitively trust an AI recommendation when it is backed by familiar examples. By presenting concrete past cases, Similar Instances align with how people naturally construct mental models based on familiar situations, as described in Kolodner's theory of case-based reasoning (Kolodner, 1992), which suggests that people interpret new situations by retrieving and adapting solutions from similar prior experiences to support decision-making. However, this pattern was not observed in the Responsive-all (RA) group, possibly because participants were not exposed to XAI in every task, indicating that the effect of Similar Instances on *Agreement* may depend on consistent reinforcement.

For participants who viewed XAI before and after AI errors, there was a sharper decline in *Agreement* with Saliency Map during the Trust Violation phase, when compared to Similar Instances. This may be due to the Saliency Map's ability to visually highlight specific areas that contributed to the AI's decision, which could have helped users identify where the AI's attention was potentially misplaced (Zeiler & Fergus, 2014). In contrast, Similar Instances cause a greater *Agreement* with AI and a corresponding lower *Decision Accuracy* during the Trust Violation phase compared to Saliency Map, which suggests that example-based XAI may have caused over-reliance on AI decisions by presenting similar examples from the training dataset. This could have created a false sense of confidence, which may have led participants to believe that the AI's past accuracy on similar tasks guaranteed correctness in the current case (Guidotti et al., 2019). This can result in participants agreeing with the AI without adequate scrutiny. This finding contrasts with prior studies that reported positive effects of Similar Instances on trust calibration and decision making (Leichtmann et al., 2023, 2024; Yang et al., 2020). One possible reason is that we presented similar examples exclusively from the target class, while those studies included examples from multiple classes, enabling users to more easily identify inconsistencies by comparing across various classes. In our setup, participants may have trusted the AI when the target resembled examples from the predicted class, while in other work, more similar instances from other classes may have highlighted potential errors. This underscores the importance of explanation fidelity and quality in effectively calibrating trust. Additionally, our prediction targets were likely more familiar to participants (e.g., common prohibited items) compared to domains like mushroom or leaf recognition, indicating that Similar Instances may be more effective in unfamiliar tasks where users have limited prior expertise. The divergent effectiveness of explanation types may also arise from task-specific demands. Object detection tasks, such as baggage screening, require spatial localization of relevant objects within cluttered images, which could make feature-based explanations that highlight the model's focus particularly helpful. In contrast, classification tasks, such as mushroom recognition, rely more on holistic pattern identification and comparison, so example-based explanations that demonstrate prototypical instances can be more effective in supporting people's classificatory reasoning. Besides, Leichtmann et al. (2023) presented both Saliency Map and Similar Instances to people at the same time, suggesting that combining XAI methods may offer complementary benefits not captured by showing a single type of explanation. Our results also showed that participants exposed to Saliency Map found the AI more familiar than those who saw Similar Instances, echoing the prior finding that feature-based explanations enhance users' understanding more effectively than example-based explanations (Wang & Yin, 2022). It may be because Saliency Map provides a mechanism, visually revealing how the AI weighs different features in its decision-making process, which more naturally aligns with human cognitive patterns of causal attribution and fosters familiarity and understanding (Zeiler & Fergus, 2014).

5.1.2. RQ2: Impact of AI error Type on trust and user verification behavior

In this study, AI errors were intentionally introduced to trigger participants' notice of them. It was expected that participants who became aware of AI errors, regardless of the type of error, would possibly seek assistance to validate their judgments about AI. Our results show that during the Trust

Violation phase, participants were more likely to request explanations (i.e., *See XAI*) and check the ground truth (i.e., *Open Baggage*) when encountering False Alarm errors, compared to the Trust Formation and Trust Repair phases. However, for Miss errors, the frequency of requesting *See XAI* and *Open Baggage* either remained similar or significantly decreased during the Trust Violation phase compared to the Trust Formation and Trust Repair phases. This difference may be explained by the nature of the explanations of errors. The XAI (explanations) for Miss errors tend to be less informative than the visual explanations provided for False Alarm. Once participants realized that explanations for Miss were not helpful, they became less likely to request XAI in those cases.

Generally, during the Trust Violation phase, False Alarm resulted in an increased frequency of *See XAI* and *Open Baggage* over Miss. This pattern may be attributed to the fact that alerts, whether accurate or not, serve as active calls to action, compelling individuals to seek additional information (Meyer et al., 2014). Additionally, this behavior is consistent with the concept of *ambiguity aversion*, a tendency to avoid options with uncertain outcomes (Ellsberg, 1961). A False Alarm could introduce uncertainty about AI's reasoning because the AI has incorrectly flagged a safe item as a threat. Participants are motivated to resolve this ambiguity by verifying the AI decision themselves. In contrast, when the AI misses an illicit object, there is less perceived uncertainty since the AI did not raise an alert, leading participants to trust their observations more readily. Even if the Miss is noticed, people may assume it was an oversight rather than a sign of unpredictable reasoning. Participants were more willing to request XAI during False Alarm errors might also be due to the expectation of visual explanations for AI's positive detection result, compared to the text explanations for the negative detection result. It also implies a potential for the visual representation of XAI to increase human-AI trust during the phase of low level of trust (i.e., Trust Violation), motivating participants to engage with the AI and be open to understanding its reasoning, even when it errs.

The group exposed to False Alarm errors also viewed the AI as more harmful and less trustworthy than the group that encountered the Miss. This difference might be because visual explanations for False Alarm errors provide participants with clearer evidence to verify AI mistakes compared to the text explanations for Miss errors. Once participants identify and understand potential causes of AI errors through XAI, their subjective perception of AI's capability decreases.

5.1.3. RQ3: Influence of explanation delivery strategy on trust and perceived trustworthiness

RQ3 addressed *how often* and *when* explanations were provided. Our findings show these significantly impacted participants' perceived trust in the AI. Participants who were continuously shown explanations reported the lowest levels of confidence and trust in the AI by the end of the task. In contrast, those who were shown explanations on demand reported the AI to be less harmful and overall trusted it more. We could attribute this to a trade-off. Participants bombarded with non-stop explanations may have experienced cognitive overload or explanation fatigue. While each explanation might clarify the AI's reasoning locally, the global effect could have been a sense of being overwhelmed or frustrated, especially in a time-sensitive task. This also highlights that there may be potential tradeoffs between local and global explanations (Bansal et al., 2021). Additionally, it is also possibile that seeing reasoning for every prediction, especially when participants noticed AI errors, made them more aware of the AI's flaws or trivial rationale which diminished their perceived dependability of the AI (Leichtmann et al., 2023). These findings resonate with a theme emerging in the literature that transparency does not equate to trust if not handled judiciously (Von Eschenbach, 2021).

5.2. Implications for supporting human appropriate trust evolution with XAI

Our findings highlight the importance of the thoughtful selection and delivery timing of XAI to advance human-AI collaboration performance while considering the impact of the type of AI errors. We found that while example-based XAI, such as Similar Instances, can enhance trust perceptions and increase agreement with AI decisions when AI results are correct, it can also lead to potentially poor decision-making during AI errors. Conversely, Saliency Map, which highlights specific regions of an image that contributed to the AI decision, might lead people to make more informed decisions during AI errors due to their ability to reveal where the AI's attention was misplaced. This suggests that the effectiveness of XAI may vary across different trust phases. Additionally, as demonstrated in our study, humans' trust in AI recovered after exposure to AI errors, returning to levels comparable to or even exceeding initial trust. This recovery was expected since the AI was deliberately designed for reliable performance. Thus, restoring human trust in AI can lead to improved overall performance. However, trust repair may not always be beneficial in real-world settings, particularly when the AI is prone to errors. This underscores the need to carefully consider when and how to implement trust-enhancing mechanisms, such as XAI, to ensure trust is appropriately calibrated during human-AI interactions.

In use cases wherein AI errors are not easily detectable by humans, AI-assisted systems could integrate explanations with confidence metrics to reduce over-reliance on AI, helping people gauge the reliability of AI predictions. Employee training on new AI technologies should not only cover its usability but also its fallibility in edge cases, emphasizing the need for independent verification and reinforcing the "trust but verify" paradigm, particularly in ambiguous or high-stake decisions. In addition, periodic reminders of AI's limitations, such as highlighting inconsistencies in decisions across similar examples, can further encourage scrutiny. Moreover, AI systems could also allow users to provide feedback on the AI's explanations to iteratively improve alignment between users' mental models and AI behaviors.

Our study also revealed that seeing AI explanations increases the frequency of human compliance with the AI, as reflected in higher agreement with AI outcomes. However, excessive exposure to XAI may overwhelm people, resulting in lower perceived trust. Balancing the frequency and timing of XAI presentations is therefore critical. Explanations should be provided for people at appropriate moments, including but not limited to when requested or during challenging tasks. For instance, systems could automatically detect difficult tasks for AI by monitoring performance indicators like prediction confidence and accuracy in similar tasks. In such cases, XAI could be delivered proactively, while users could retain the option to request explanations for less difficult tasks. This approach could help maintain human trust while reducing cognitive overload, which is particularly critical for high-stakes and time-sensitive tasks. As such, designing adaptive AI systems that incorporate various XAI techniques and delivery strategies may achieve complementary effects, effectively supporting people at various stages of trust development. Furthermore, understanding long-term trust dynamics requires considering how peoples' trust in AI evolves with their personal experiences and the intent behind interactions over time. Mechanisms that allow people to provide feedback on their trust and confidence in AI could enable the system to tailor explanation strategies to better meet individual needs, promoting appropriate trust. This is particularly valuable in situations where AI error is not obvious, and human trust in AI is hard to predict.

As observed from our study, different factors have varied effects on diverse trust-related measures, including *Agreement* with AI, *Open Baggage*, *See XAI*, and perceived trust. Even the same behavior, such as the option of *See XAI*, can signal different trends in trust growth or decline depending on the level of human trust in AI at a specific time. These variabilities highlight the importance of viewing trust as a multifaceted construct when evaluating its dynamics in human-AI collaboration, encouraging practitioners to approach human-AI trust in decision-making from multiple dimensions.

5.3. Limitations and future work

This research study used a synthetic test environment (STE) and general population participants. Thus, conclusions and findings regarding trust calibration, the effectiveness of explanation types, and responses to AI errors should be interpreted with caution. We recognize that novices may differ from domain experts in how they perceive and respond to AI systems and their explanations (Ehsan et al., 2024). Experts, due to extensive training and professional experience, may exhibit higher skepticism toward AI recommendations (Gaube et al., 2021; Nourani et al., 2020a), have established mental models of what constitutes meaningful explanations, and react differently to error types, particularly missed detections which carry significant operational consequences in real-world security screening. Future research should validate these findings with domain experts in realistic operational settings. Comparative studies involving novices and experts could help in understanding how professional experience and domain knowledge mediate trust dynamics, AI error perception, and explanation effectiveness in AI-assisted security screening. To address this limitation of our work and to validate existing findings, we have initiated an ongoing field study involving security officers from the Transportation Security Administration (TSA).

Our STE also exhibits important limitations that could impact ecological validity. In real-world airport security operations, baggage without threats significantly outnumbers threat-containing bags. Our test environment did not replicate this natural imbalance, potentially influencing participants' vigilance, decision-making patterns, and trust responses. Compared to domains such as medical image diagnosis or pedestrian detection, baggage screening features a lower true positive rate but carries far more severe consequences for missed detections. These task-specific factors, such as the rarity and severity of errors, perceived risks, and urgency, can significantly affect operators' behavior, influencing how they weigh false positives versus false negatives, how much they rely on AI, and how their trust in AI develops over time. To improve generalizability, future work should incorporate more realistic threat prevalence rates, examine various detection contexts with different risk levels, and explore how these task-specific characteristics interact with explanation methods and strategies to shape human trust and decision behavior.

Additionally, participants began with 12 consecutive AI-correct tasks to initiate a baseline level of trust. However, this approach risks inducing an overly strong positive first impression of the AI and promoting automation bias (Desai et al., 2013; Nourani et al., 2022), leading participants to over-rely on the AI even when its performance is poor (e.g., during the Trust Violation phase). We also used simplified task scenarios designed to induce trust violation by presenting participants with AI errors that were intentionally easy for people to visually verify. This approach might not fully capture the complexities of real-world situations. In practice, AI models may be more likely to make mistakes in challenging tasks that are less obvious for people to notice. Errors in such complex scenarios may be less apparent but potentially more acceptable to people. We also manipulated a specific pattern of three consecutive errors to cause trust violations. Yet, variations in error frequency (i.e., system error rate), types of errors, or patterns (e.g., single or mixed error types) could have different effects on human trust evolution, even in the presence of XAI. Future work should explore varying AI performance for trust formation to minimize potential biases.

The meaningfulness of the XAI (Nourani et al., 2019) could also influence human-AI trust in distinct ways. For example, we also did not address the challenge of effectively explaining the negative results of baggage screening tasks, i.e., AI detected no illicit object in baggage. Future research could explore how different XAI methods can be used to improve the explanation of such cases.

And finally, future work could extend to long-term trust evolution (e.g., include more task sessions) and incorporate a broader range of XAI techniques, XAI communication strategies, and task categories. This would offer a more comprehensive understanding of how trust evolves in the presence of XAI. Moreover, advanced statistical models, such as Structural Equation Modeling (SEM), can be employed to combine people's various behaviours or trust-related perception scores as a composite measure, enabling a more integrated analysis for the underlying human-AI trust latent.

6. Conclusion

In this study, we explored the impacts of XAI on human trust and human-AI collaborative performance over time, particularly focusing on how different types of Explainable AI (XAI) techniques, AI errors, and strategies for delivering explanations impact user trust and decision-making in a high-stakes AI-assisted decision-making scenario (i.e., baggage screening). Our study measured human trust in AI from multiple perspectives (e.g., human behavior and perception) and unveiled that human trust was influenced by various interacting factors such as the type of XAI used and the AI correctness. Specifically, we observed that Similar Instances and Saliency Map can lead to a comparable level of human trust in AI, as well as decision accuracy, when AI makes correct decisions. Yet, when AI makes mistakes, Similar Instances could result in human over-reliance on AI while Saliency Map promotes a more cautious approach and guides users to make more informed decisions, though they may result in decreased agreement with AI decisions. In addition, the strategy of providing explanations—whether consistently, based on task difficulty, or upon request—significantly affects human perceived trust, with on-demand explanations leading to higher perceived trust than constant exposure. Besides, the type of AI error that people experience during the human-AI collaboration can remarkably affect their desire to know explanations about the AI outcome as well as their confidence in making decisions. In conclusion, the impact of XAI on human-AI trust and collaborative performance over time is complex and varies with changes in AI performance, different strategies for presenting XAI, and varying types of AI errors. These insights have crucial implications for designing human-AI collaboration systems, emphasizing the need to tailor XAI approaches based on human needs, AI performance, and error contexts if applicable. To enhance long-term trust and collaboration, systems should consider the multifaceted nature of trust, providing appropriate explanations that are needed by people for specific tasks. By understanding and managing the trust evolution, designers can create human-AI collaboration systems that not only build initial trust but also sustain and adapt it over time, ultimately leading to more effective and reliable human-AI collaboration.

Notes

1. https://osf.io/6mbk4/
2. https://github.com/MeioJane/SIXray
3. https://osf.io/6mbk4/

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-08-00. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

ORCID

Yixuan Wang http://orcid.org/0000-0003-0195-1193
Jieqiong Zhao http://orcid.org/0000-0002-4303-7722
Yang Ba http://orcid.org/0009-0000-8237-5962
Michelle V. Mancenido http://orcid.org/0000-0002-3000-8922
Erin K. Chiou http://orcid.org/0000-0002-7201-8483
Ross Maciejewski http://orcid.org/0000-0001-8803-6355

References

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

Baer, M. D., & Colquitt, J. A. (2018). Why do people trust?: Moving toward a more comprehensive consideration of the antecedents of trust. In *The Routledge Companion to Trust* (pp. 163–182).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 81:1–81:16). ACM.

Baughan, A., Wang, X., Liu, A., Mercurio, A., Chen, J., & Ma, X. (2023). A mixed-methods approach to understanding user trust after voice assistant failures. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 7:1–7:16). ACM.

Bliss, J. P., & Fallon, C. K. (2006). Active warnings: False alarms. In M. S. Wogalter (Ed.), *Handbook of warnings* (1 ed., pp. 231–242). CRC Press.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the international conference on intelligent user interfaces* (pp. 454–464). ACM.

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal*, *40*(5), 685–706. https://doi.org/10.1016/j.emj.2022.06.001

Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *Proceedings of the international conference on intelligent user interfaces* (pp. 258–262). ACM.

Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2020). Some cues are more equal than others: Cue plausibility for false alarms in baggage screening. *Applied Ergonomics*, *82*, 102916. https://doi.org/10.1016/j.apergo.2019.102916

Chen, C., & Sundar, S. S. (2023). Is this AI trained on credible data? The effects of labeling quality and performance bias on user trust. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 816:1–816:11). ACM.

Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human–Computer Interaction*, *7*(CSCW2), 1–32. https://doi.org/10.1145/3610219

Cheng, H.-F., Wang, R., Zhang, Z., O'connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 559:1–559:12). ACM.

Chien, S.-Y., Lewis, M., Sycara, K., Kumru, A., & Liu, J.-S. (2020). Influence of culture, transparency, trust, and degree of automation on automation use. *IEEE Transactions on Human-Machine Systems*, *50*(3), 205–214. https://doi.org/10.1109/THMS.2019.2931755

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, *81*, 59–83. https://doi.org/10.1016/j.inffus.2021.11.003

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *Proceedings of the international conference on intelligent user interfaces* (pp. 307–317). ACM.

Collins, L. M., Dziak, J. J., Kugler, K. C., & Trail, J. B. (2014). Factorial experiments: Efficient tools for evaluation of intervention components. *American Journal of Preventive Medicine*, *47*(4), 498–504. https://doi.org/10.1016/j.amepre.2014.06.021

Congress, U. (2020). *Congressional record*. https://www.congress.gov/congressional-record/volume-166/issue-43/house-section/article/H1477-6/

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496. https://doi.org/10.1007/s11257-008-9051-3

Dasgupta, A., Burrows, S., Han, K., & Rasch, P. J. (2017). Empirical analysis of the subjective impressions and objective measures of domain scientists' visual analytic judgments. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 1193–1204).

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *Proceedings of the ACM/IEEE International conference on human–robot interaction* (pp. 251–258). IEEE.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, *162*, 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv Preprint arXiv:1710.00794.

Eckhardt, S., Kühl, N., Dolata, M., & Schwabe, G. (2024). A survey of ai reliance. arXiv Preprint arXiv:2408.03948.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2024). The who in XAI: How AI background shapes perceptions of ai explanations. In *Proceedings of the ACM conference on human factors in computing systems, CHI '24*. ACM.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669. https://doi.org/10.2307/1884324

Esterwood, C., & Robert, L. P. (2021). Do you still trust me? Human–robot trust repair strategies. In *Proceedings of the IEEE International conference on robot & human interactive communication, RO-MAN* (pp. 183–188). IEEE.

Ferraro, J. C., & Mouloua, M. (2021). Automated alert failures and their impact on operator performance and trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 1286–1290. https://doi.org/10.1177/1071181321651290

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine*, *4*(1), 31. https://doi.org/10.1038/s41746-021-00385-9

Gerlings, J., Jensen, M. S., & Shollo, A. (2022). Explainable AI, but explainable to whom? In *An exploratory case study of xAI in healthcare* (pp. 169–198). Springer.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2020). Black box explanation by learning image exemplars in the latent feature space. In *Proceedings of the European conference on machine learning and knowledge discovery in databases* (pp. 189–205). Springer.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., & Hsiung, C.-P. (2019). Positive bias in the 'trust in automated systems survey'? An examination of the Jian et al. (2000) scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 217–221. https://doi.org/10.1177/1071181319631201

Hald, K., Weitz, K., André, E., & Rehm, M. (2021). 'An error occurred!' - Trust repair with virtual robot using levels of mistake explanation. In *Proceedings of the International conference on human–agent interaction* (pp. 218–226). ACM.

Hartnett, G. S., Held, B., & McKay, S. (2022). Airline security through artificial intelligence: How the transportation security administration can use machine learning to improve the electronic baggage screening program. *Policy Commons*. https://www.jstor.org/stable/resrep40391

Hättenschwiler, N., Sterchi, Y., Mendes, M., & Schwaninger, A. (2018). Automation in airport security x-ray screening of cabin baggage: Examining benefits and possible implementations of automated explosives detection. *Applied Ergonomics*, *72*, 58–68. https://doi.org/10.1016/j.apergo.2018.05.003

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, *28*(1), 84–88. https://doi.org/10.1109/MIS.2013.24

Holliday, D., Wilson, S., & Stumpf, S. (2016). User trust in intelligent systems: A journey over time. In *Proceedings of the international conference on intelligent user interfaces* (pp. 164–168). ACM.

Hu, W.-L., Akash, K., Reid, T., & Jain, N. (2019). Computational modeling of the dynamics of human trust during human–machine interactions. *IEEE Transactions on Human-Machine Systems*, *49*(6), 485–497. https://doi.org/10.1109/THMS.2018.2874188

Huegli, D., Chavaillaz, A., Sauer, J., & Schwaninger, A. (2025). Effects of false alarms and miscues of decision support systems on human–machine system performance: A study with airport security screeners. *Ergonomics*, 1–16. https://doi.org/10.1080/00140139.2025.2453546

Humer, C., Hinterreiter, A., Leichtmann, B., Mara, M., & Streit, M. (2022). Comparing effects of attribution-based, example-based, and feature-based explanation methods on AI-assisted decision-making. *OSF Preprints*. https://doi.org/10.31219/osf.io/h6dwz

Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021). Explainable artificial intelligence approaches: A survey. arXiv Preprint arXiv:2101.09429.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 624–635). ACM.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kahr, P. K., Rooks, G., Snijders, C., & Willemsen, M. C. (2024). The trust recovery journey. The effect of timing of errors on the willingness to follow AI advice. In *Proceedings of the international conference on intelligent user interfaces* (pp. 609–622). ACM.

Kahr, P. K., Rooks, G., Willemsen, M. C., & Snijders, C. C. (2023). It seems smart, but it acts stupid: Development of trust in AI advice in a repeated legal decision-making task. In *Proceedings of the international conference on intelligent user interfaces* (pp. 528–539). ACM.

Kay, M., Patel, S. N., & Kientz, J. A. (2015). How good is 85%? A survey tool to connect classifier evaluation to acceptability of accuracy. In *Proceedings of the 33rd Annual ACM Conference on human factors in computing systems* (pp. 347–356).

Keller, E. (2019). *One in five TSA screeners quits within six months*. https://www.govexec.com/pay-benefits/2019/04/one-four-tsa-screeners-quits-within-six-months/156045/

Kim, S. S., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). "Help me help the ai": Understanding how explainability can support human–AI interaction. In *Proceedings of the ACM conference on human factors in computing systems*. ACM.

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*, 101595. https://doi.org/10.1016/j.tele.2021.101595

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, *6*(1), 3–34. https://doi.org/10.1007/BF00155578

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. In *Proceedings of the IEEE symposium on visual languages and human centric computing* (pp. 3–10). IEEE.

Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., & Ziegler, J. (2019). Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 487:1–487:12). ACM.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 29–38). ACM.

Latscha, M., Schwaninger, A., Sauer, J., & Sterchi, Y. (2024). Performance of x-ray baggage screeners in different work environments: Comparing remote and local cabin baggage screening. *International Journal of Industrial Ergonomics*, *102*, 103598. https://doi.org/10.1016/j.ergon.2024.103598

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., & Mara, M. (2024). Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human–Computer Interaction*, *40*(17), 4787–4804. https://doi.org/10.1080/10447318.2023.2221605

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Lewicki, R. J., & Bunker, B. B. (1996). Developing and maintaining trust in work relationships. In *Trust in organizations: Frontiers of theory and research, chapter 7* (pp. 114–139). SAGE Publications.

Liang, K. J., Sigman, J. B., Spell, G. P., Strellis, D., Chang, W., Liu, F., Mehta, T., & Carin, L. (2019). Toward automatic threat recognition for airport x-ray baggage screening with deep convolutional object detection. arXiv Preprint arXiv:1912.06329.

Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human–Computer Interaction*, *5*(CSCW2), 1–45. https://doi.org/10.1145/3479552

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Marinaccio, K., Kohn, S., Parasuraman, R., & de Visser, E. (2015). A framework for rebuilding trust in social automation across health-care domains. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *4*(1), 201–205. https://doi.org/10.1177/2327857915041036

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Mayr, E., Hynek, N., Salisu, S., & Windhager, F. (2019). Trust in information visualization. In *Proceedings of the Eurographics Conference on visualization workshop on trustworthy visualization* (pp. 25–29). The Eurographics Association.s

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *The Academy of Management Review*, *23*(3), 473–490. https://doi.org/10.2307/259290

Meyer, J., Wiczorek, R., & Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors*, *56*(5), 840–849. https://doi.org/10.1177/0018720813512865

Miao, C., Xie, L., Wan, F., Su, C., Liu, H., Jiao, J., & Ye, Q. (2019). SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2119–2128). IEEE.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, *31*(3), 175–178. https://doi.org/10.1016/S0169-8141(02)00194-4

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect xai on human–AI decision-making. *Proceedings of the ACM on Human–Computer Interaction*, *8*(CSCW1), 1–39. arXiv: 23073566. https://doi.org/10.1145/3641022

Mundhenk, T. N., Chen, B. Y., & Friedland, G. (2020). Efficient saliency maps for explainable AI. arXiv Preprint arXiv:1911.11293.

Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI conference on human computation and crowdsourcing* (Vol. 7, pp. 97–105). AAAI.

Nourani, M., King, J., & Ragan, E. (2020a). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*(1), 112–121. https://doi.org/10.1609/hcomp.v8i1.7469

Nourani, M., King, J. T., & Ragan, E. D. (2020b). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI conference on human computation and crowdsourcing* (Vol. 8, pp. 112–121). AAAI.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E. D., & Gogate, V. (2022). On the importance of user backgrounds and impressions: Lessons learned from interactive AI applications. *ACM Transactions on Interactive Intelligent Systems*, *12*(4), 1–29. https://doi.org/10.1145/3531066

Ooge, J., Kato, S., & Verbert, K. (2022). Explaining recommendations in e-learning: Effects on adolescents' trust. In *Proceedings of the international conference on intelligent user interfaces* (pp. 93–105). ACM.

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer–Human Interaction*, *29*(4), 1–33. https://doi.org/10.1145/3495013

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Pareek, S., Velloso, E., & Goncalves, J. (2024). Trust development and repair in AI-assisted decision-making during complementary expertise. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 546–561). ACM.

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, *286*(3), 800–809. https://doi.org/10.1148/radiol.2017171920

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., & Saenko, K. (2021). Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11438–11447). IEEE.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 273:1–273:52). ACM.

Prinster, D., Mahmood, A., Saria, S., Jeudy, J., Lin, C. T., Yi, P. H., Huang, C.-M., Moy, L., & Wolfe, S. (2024). Care to explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology*, *313*(2), e233261. https://doi.org/10.1148/radiol.233261

Quinn, D. B., Pak, R., & de Visser, E. J. (2017). Testing the efficacy of human–human trust repair strategies with machines. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1794–1798. https://doi.org/10.1177/1541931213601930

Rao, Q., & Frtunikj, J. (2018). Deep learning for self-driving cars: Chances and challenges. In *Proceedings of the 1st International workshop on software engineering for* AI in *autonomous systems* (pp. 35–38).

Rebensky, S., Carmody, K., Ficke, C., Nguyen, D., Carroll, M., Wildman, J., & Thayer, A. (2021). Whoops! something went wrong: Errors, trust, and trust repair strategies in human agent teaming. In *Proceedings of the international conference on human–computer interaction* (pp. 95–106). Springer.

Reich, T., Kaju, A., & Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *33*(2), 285–302. https://doi.org/10.1002/jcpy.1313

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. In *Proceedings of the international conference on social robotics* (pp. 574–583). Springer.

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2025). How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (tram). *Computers in Human Behavior*, *170*, 108671. https://doi.org/10.1016/j.chb.2025.108671

Schwaninger, A., Hardmeler, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems Magazine*, *20*(6), 29–35. https://doi.org/10.1109/MAES.2005.1412124

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science*, *16*(5), 264–268. https://doi.org/10.1111/j.1467-8721.2007.00517.x

Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., & Perer, A. (2023). Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 754:1–754:18). ACM.

Szymanski, M., Verbert, K., & Vanden Abeele, V. (2022). Designing and evaluating explainable AI for non-AI experts: Challenges and opportunities. In *Proceedings of the ACM conference on recommender systems* (pp. 735–736). ACM.

Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., & Bernstein, A. (2021). Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the ACM conference on user modeling, adaptation and personalization* (pp. 77–87). ACM.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, *1*(4), 100049. https://doi.org/10.1016/j.patter.2020.100049

TSA. (2024). *Checkpoint requirements and planning guide (crpg)*. TSA.gov.

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? a survey of empirical methodologies. *Proceedings of the ACM on Human–Computer Interaction*, *5*(CSCW2), 1–39. https://doi.org/10.1145/3476068

Verma, N. (2021). *Deep image search - AI-based image search engine*. https://github.com/TechyNilesh/DeepImageSearch

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887. https://doi.org/10.48550/arXiv.1711.00399

Wang, X., Lu, Z., & Yin, M. (2022). Will you accept the AI recommendation? Predicting human behavior in AI-assisted decision making. In *Proceedings of the ACM web conference* (pp. 1697–1708). ACM.

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the international conference on intelligent user interfaces* (pp. 318–328). ACM.

Wang, X., & Yin, M. (2022). Effects of explanations in AI-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems*, *12*(4), 1–36. https://doi.org/10.1145/3519266

Wang, X., & Yin, M. (2023). Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making. In *Proceedings of the ACM conference on human factors in computing systems*. ACM.

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/bdm.414

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology: CB*, *20*(2), 121–124. https://doi.org/10.1016/j.cub.2009.11.066

Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell me about yourself: Using an AI-powered Chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction, 27*(3), 15:1–15:37. https://doi.org/10.1145/3381804

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the international conference on intelligent user interfaces* (pp. 189–201). ACM.

Yang, X. J., Schemanske, C., & Searle, C. (2023). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, *65*(5), 862–878. https://doi.org/10.1177/00187208211034716

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 279:1–279:12). ACM.

Yu, K., Berkovsky, S., Conway, D., Taib, R., Zhou, J., & Chen, F. (2016). Trust and reliance based on system accuracy. In *Proceedings of the ACM conference on user modeling, adaptation and personalization* (pp. 223–227). ACM.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the international conference on intelligent user interfaces* (pp. 307–317). ACM.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision, ECCV 2014* (pp. 818–833). Springer.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 295–305). ACM.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition, CVPR 2016* (pp. 2921–2929). IEEE.

Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2022). Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 1161–1171. https://doi.org/10.1109/TVCG.2021.3114864

About the authors

Yixuan Wang is a PhD student in Computer Science at Arizona State University. She received the MS from Northeastern University, MA, in 2020. Her research interests include visual analytics, explainable AI and AI uncertainty, interactive machine learning, human-computer interaction, and human-AI teaming.

Jieqiong Zhao received the MS from Tufts University in 2013 and the PhD from Purdue University in 2020. She is an assistant professor with the Department of Computer Science at Augusta University. She was a postdoc at Arizona State University. Her research focuses on data visualization, HCI, and human-AI teaming.

Yang Ba is a PhD student in Data Science, Analytics, and Engineering at Arizona State University. His research focuses on machine learning and natural language processing, specializing in synthetic data generation, evaluation, and using synthetic data to analyze model generalizability and uncertainty.

Michelle V. Mancenido is an associate professor in the School of Mathematical and Natural Sciences. Mancenido's research focuses on the design and analysis of statistical experiments in engineering, scientific, and industrial applications. She is an advocate of well-designed experiments as the key to robust scientific conclusions and efficient industrial processes.

Erin K. Chiou is an associate professor of human systems engineering at The Polytechnic School at Arizona State University. Chiou directs the Automation Design Advancing People and Technology (ADAPT) Laboratory, where graduate and undergraduate students conduct laboratory and field research studies on human-automation interaction and job design in sociotechnical systems.

Ross Maciejewski is a Professor at Arizona State University and Director of the School of Computing and Augmented Intelligence. His primary research interests are in the areas of visualization and explainable AI.